Evolutionary analyses of non-genealogical bonds produced by introgressive descent

Eric Bapteste*,†, Philippe Lopez‡, Frédéric Bouchard§, Fernando Baquero*, James O. McInerney¶, and Richard M. Burian**

*Unité Mixte de Recherche 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, 75005 Paris, France; ‡Département de Philosophie, Université de Montréal, Montréal, QC, Canada H3C 3J7; §Department of Microbiology, Ramón y Cajal University Hospital (IRYCIS, CIBERESP), 28034 Madrid, Spain; ¶Molecular Evolution and Bioinformatics Unit, Department of Biology, National University of Ireland Maynooth, County Kildare, Ireland; and **Department of Philosophy, Virginia Tech, Blacksburg, VA 24061

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved September 24, 2012 (received for review April 20, 2012)

All evolutionary biologists are familiar with evolutionary units that evolve by vertical descent in a tree-like fashion in single lineages. However, many other kinds of processes contribute to evolutionary diversity. In vertical descent, the genetic material of a particular evolutionary unit is propagated by replication inside its own lineage. In what we call introgressive descent, the genetic material of a particular evolutionary unit propagates into different host structures and is replicated within these host structures. Thus, introgressive descent generates a variety of evolutionary units and leaves recognizable patterns in resemblance networks. We characterize six kinds of evolutionary units, of which five involve mosaic lineages generated by introgressive descent. To facilitate detection of these units in resemblance networks, we introduce terminology based on two notions, P3s (subgraphs of three nodes: A, B, and C) and mosaic P3s, and suggest an apparatus for systematic detection of introgressive descent. Mosaic P3s correspond to a distinct type of evolutionary bond that is orthogonal to the bonds of kinship and genealogy usually examined by evolutionary biologists. We argue that recognition of these evolutionary bonds stimulates radical rethinking of key questions in evolutionary biology (e.g., the relations among evolutionary players in very early phases of evolutionary history, the origin and emergence of novelties, and the production of new lineages). This line of research will expand the study of biological complexity beyond the usual genealogical bonds, revealing additional sources of biodiversity. It provides an important step to a more realistic pluralist treatment of evolutionary complexity.

*Author contributions: E.B., F. Bouchard, F. Baquero, and R.M.B. designed research; P.L. performed research; J.O.M. contributed new reagents/analytic tools; E.B. and P.L. analyzed data; and E.B., P.L., F. Bouchard, F. Baquero, J.O.M., and R.M.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: eric.bapteste@env.jussieu.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1206541109/-/DCSupplemental.
generating genetic combinations have produced a variety of evolutionary outcomes at different hierarchical levels (26). Examples include domain sharing between gene families (27), transfer of adaptive genes in prokaryotic genomes (28–32), pangenomes (33), and sharing of transposases (34), integron gene cassettes (29), plasmids (35), and phages (28, 31) within genetic exchange communities (36); bacterial consortia, such as Chlorochromatium aggregatum, with partners undergoing synchronized separate cellular divisions (37); and endosymbiotic gene transfer (38, 39).

In cases of symbiosis, mutualistic, commensal, and even parasitic relationships, gene exchange is not a necessary condition for the formation of higher-order entities that are composed of separate units with their own genomes. The contributing entities can profit from the combined resources made possible by interactions between the products encoded by the genes of the partners and can also yield an entity that is subject to selection in its own right. For instance, biofilms; colonial organisms [Volvox (40), sponges, Portuguese Man-O’-War, and the aggregates and slugs of Dictyostelium discoideum (41)]; multicellular eukaryotes, insect hosts, and the Wolbachia that determine their sex or other traits (42, 43); lichens (44, 45), herds and packs of social animals, communally organized (quasisocial and eusocial) social insects; and commensal and symbiotic gut microbes of insects and vertebrates together with their hosts are all excellent candidates to count as higher-order entities (or collective reproducers) (18). The genome of such collective reproducing entities can pro

What we call introgressive descent occurs precisely when the genetic material of a particular entity propagates into different host structure(s) and then is propagated within or by the resulting unit(s). Examples include a transposon inserted into a series of different plasmids, a plasmid in different bacterial clones, a clone in different microbiomes, the mitochondrial genes present in a eukaryotic cell (regardless of whether those genes have been transferred into the nuclear genome), and the commensal combination of an alga and a fungus in a lichen that is propagated by vegetative reproduction or diaspores (44, 45, 50). The typical biological outcomes of these interlineage and interlevel assortments, namely the mosaic objects, and the multilocus coalitions of genetic partners involved in these processes can be stabilized and selected, becoming important evolutionary players in their own right (46). Therefore, introgressive descent generates non-genealogical bonds between biological objects, producing a reticulate evolutionary framework.

To account for the origins and features of these objects, we propose that, in

Mergers and Clubs as Relevant Evolutionary Units. Members of monophyletic groups, evolving by clonal division and allowing for continuing mutational diversification in members of clonal complexes, characteristically share genes that trace back to a single locus in a single individual (in fact, the same locus in a single genome of a last common ancestor). We call such genes coalescent orthologs to distinguish them from shared genes originating from different processes. Indeed, many genetic similarities between biological objects are not caused by vertical descent, where the genetic material of a particular entity is propagated by replication inside its own lineage. For instance, adaptive lateral genetic transfer between genomes of entities from different lineages that share the same environment or lifestyle (29, 32, 46) indicates additional (non-vertical) mechanisms for the integration of genetic material into one host. Hence, another type of descent is fundamental to the re-construction of an accurate evolutionary picture of the evolutionary units.
addition to single lineages resulting from the splitting processes of vertical descent, evolutionists should formally recognize a range of mosaic evolutionary units produced by introgressive descent. This range has two extremes. First, there are mergers. Mergers arise when two or more components, not hitherto coexisting within the same unit, are brought together, and these components are subsequently replicated or propagated within or by a new single corporate body (9). Often, component parts of mergers do not trace back to a single locus (or set of loci) in a single last common ancestor. Mergers exist at multiple levels of biological organization [molecular (27, 51), genomic (25, 52–54), and organismal (39, 55, 56)] and do not all subsume the same genetic consequences. Fused genes conferring drug resistance (35), new viral genomes (49), lineages created from symbioses (39, 56), and Russian dolls of mobile genetic elements (52, 53) are among the best known examples of mergers. The offspring of sexual reproduction are also obligate mergers, because their parts come from distinct—although closely related—sources (two parents instead of one last common ancestor). Many mergers bring together elements that were capable of independent replication before and can replicate only as part of a larger whole after their union (19); in such cases, they present typical signs of evolutionary transitions.

Second, there are multilineage clubs. Members of these clubs form coalitions of entities that replicate in separate events and exploit some common genetic material that does not trace back to a single locus in a single last common ancestor of all of the members (26, 29, 31, 32, 57, 58). Multispecies biofilms (59), environmental coalitions of cells and mobile genetic elements like those elements of marine cyanophages and cyanobacteria (28), and genetic exchange communities in gut microbiomes (31, 60, 61) provide examples of such multilineage clubs. These assortments may result in evolutionary transitions if the club exhibits some form of reproduction in its own right.

Some independently reproducing components of a larger whole will also fall between these two extreme poles that are produced by introgressive descent. Thus, the mycobionts and photobionts of most lichens may reproduce independently (although in such cases, the offspring of the mycobiont must find and incorporate an appropriate photobiont to be lichenized again), but they may also reproduce by vegetative reproduction or diaspores; therefore, they may be treated as facultative mergers (44, 45, 50). In contrast, the mycobionts of some populations of lichens seem to have lost the power of independent reproduction; such lichens are (obligate) mergers for their components that cannot reproduce independently (62). Consequently, empirical evidence regarding reproduction, maintenance mechanisms, integration, and fitness of each proposed merger (or club) is required for a detailed evaluation of why particular genetic assortments (or coalitions) based on the sharing of genetic material count (or not) as bona fide evolutionary units or are on a path to an evolutionary transition (SI Text, section 2 and Fig. S1).

In fact, when embracing the common definition of lineages (where groups of closely related entities belong to the same lineage by contrast to different lineages, which refer to groups of more distantly related entities) and the common definition of levels of biological organization (with cells and mobile genetic elements belonging to different levels), we propose to distinguish no fewer than five main classes of candidate evolutionary units. These units are (i) intralineage mergers, (ii) interlineage mergers, (iii) interlevel mergers, (iv) multilineage clubs, and (v) multilevel clubs, depending on whether the genetic material shared by introgressive descent comes from a single lineage and level of biological organization or more (SI Text, sections 1 and 2).

Examination of the importance of such units should broaden (and may challenge) traditional descriptions of evolutionary history, which are still largely focused on single lineages with evolution that can be modeled by a tree. We must, therefore, think about methodological innovations to deal with these additional interactors, which can include the use of directed or undirected cyclical graphs known as networks and the use of a simple graph-based terminology.

Tracking Non-genealogical Bonds in Evolutionary Networks. Networks, consisting of nodes connected by edges, are a natural way to capture specific patterns resulting from the distribution of genetic material from more than one source (36, 47–49). These graphs can represent genetic diversity at different levels of biological organization. For instance, gene networks represent sequences by nodes, and these nodes are connected by edges when they manifest significant similarity (63). Genome networks represent genomes as nodes, and these nodes are connected by edges when they share common features (e.g., the same sequence or the same gene family) (47–49).

In genome networks, monophyletic groups will generally produce cliques (Figs. 1 and 2 and Table 1) (i.e., subgraphs in which all nodes are directly connected to one another), because all entities under study share some coalescent orthologs. However, when the similarity of characters decreases under a given threshold through evolution, a different
Fig. 2. Patterns with evolutionary significance in resemblance networks. Each symbol indicates an entity (node) from a distinct level of biological organization. Similarly colored edges indicate vertically inherited shared characters. Occurrences in our test dataset at >50% identity are quantified when available. (A) Clique (here, a triangle) capturing a homology relationship between A, B, and C. (B) P3 occurring when a homologous character evolved beyond recognition between A and C. (C) M-P3 indirectly connecting two entities through a third one by different (pink and green) shared characters. (D) Multilevel M-P3 indicating multilevel evolutionary units. (E) Polarized M-P3 showing B as a merger or as a fissioning unit. (F) Pn with the distinctly related parts from a merger entity A. (H) Multilevel M-P3s. (Left) Ancient core characters mask a recent combination of characters in B. (Center) Real numbers of shared gene families between domains of life. (Right) Aggregation of three M-P3s looking like a clique. (I) Multilevel cliques.

A pattern is produced: some edges disappear, and cliques are replaced by intransitive chains, with adjacent objects of the chain presenting similarity up to a certain threshold (Fig. 2B). In agreement with the terminology of graph theory, we call such a subgraph of three nodes (A, B, and C) a P3 (64), where A is linked to B, B is linked to C, and A is not linked to C. This concept can be easily extended to the case where A, B, and C are not nodes but instead, cliques; in graph theory, B is called a minimal clique separator (65).

By contrast, we call mosaic-P3 (M-P3) a P3, in which two entities, A and C, are indirectly connected through a third entity, B, by one or more characters that are not coalescent orthologs (Fig. 2C). Such an M-P3 unites at least two distinct related and/or unrelated lineages through a third entity acting as an intermediate binder. By definition, this structure is beyond the reach of a single-tree analysis; A and C cannot be compared directly, because they lack homology for the traits under study. The relationship between A and C is not an intrinsic property of these two objects, and it is distinct from homology. Consequently, such M-P3s offer non-geenalogical bonds to detect multilinage clubs (when all nodes of the M-P3 represent entities from different lineages but at the same level of biological organization) or multilevel clubs (when some of its nodes represent entities from different levels of biological organization; e.g., cellular chromosomes, phages, and plasmids) (Fig. 2D). Moreover, when polarized, M-P3s can be used to detect mergers (Fig. 2E) when the binder receives genetic contributions from two sources (ex pluribus unum), or M-P3s can be used to detect that a fissioning entity has contributed materially distinct objects (ex unibus pluram) (66). In both mergers and contributions to separate entities, the involved entities may belong to the same level or to different levels of biological organization.

We define Pn, when n entities can be arranged, as a chain of n-2 P3s (Fig. 2F). Importantly, a Pn can also detect mosaic units, when entities at its extremities are distinct parts of the same entity (Fig. 2G) (e.g., when the terminal nodes in a gene network are two genes present in the same organism but acquired from distinct sources).

Such simple patterns of the connections can facilitate the study of introgressive descent in networks. As a quick proof of concept, we assembled and BLASTed all-against-all, a dataset of 336,402 cellular protein sequences, from the complete genomes of 54 Archaea bacteria, 70 Eubacteria, and 7 Eukaryotes sampled all over the web of life (the taxa are listed in SI Text, section 3) and 228,042 mobile genetic element protein sequences, comprising all viral and plasmid sequences available from the National Center for Biotechnology Information as of May of 2011. These sequences are available in the download section at www.evol-net.fr. We built gene networks (www.evol-net.fr) by connecting two sequences if they shared a BLAST hit displaying more than a given percentage identity (e.g., 50%, 70%, 90%, and 99%) and considered edges corresponding to a BLAST hit covering more than 80% of both sequences as sequence-homologous. In this case, we observed 6,477 Pn patterns in our gene network, with distantly connected genes from the same homologous family in eukaryotes: one acquired from an archaebacterial ancestor, and the other acquired from a bacterial endosymbiont (mitochondria or chloroplast). Many of these Pn were tracking the same ancient event of endosymbiotic transfer.

Although M-P3s can be characterized in terms of graph theory, their detection can be complex. For instance, M-P3 patterns can be missed when either characters assumed to be coalescent orthologs are not. This situation can occur for gene families with significant amounts of in and out paralogy (67) or in the extreme case of nearly identical replacement of genetic material by sequence-homologous copies. Finally, cliques with unrelated entities (Fig. 2I) also deserve particular consideration, because they are not united by vertical descent. Their topology suggests the sharing of genetic material in multilevel clubs.

Formally naming these P3s (and cliques) is a first step for implementing their systematic detection to better track evolutionary transitions and evolutionary units using both genealogical and non-geenalogical bonds. Typically, in our real dataset, no single tree can analyze all of the connected sequences in the gene network, because no single clique with more than four sequences entirely covers a connected component unifying sequences with significant similarities (Fig. 1 and Table 1). Only a fraction of the sequences in a gene network included in such cliques (counted using maximal clique enumerator) (68) are amenable to classic phylogenetic analysis; 11.5–36.3% of the sequences are present in P3, meaning that their relationships of homology are also too distant to be accounted for by a single tree. In addition, a fair proportion of sequences (from 3.8% to 28.9%) belongs to M-P3 and multilevel P3 (up to 11.4%) subgraphs, further hinting at phenomena of introgressive descent (Table 1). Likewise, although numerous sequences belong to triangles connected by homology edges, suggesting that their similarity...
results from vertical descent, in a vast majority, these triangles contain sequences from genomes from distinct levels of organization, indicating important amounts of genetic sharing between unrelated entities. Moreover, depending on the threshold retained to construct the gene network, an additional 4.7–27.4% of triangles present in the network would rather be explained by the introgressive sharing of unrelated (or extremely divergent) fragments of DNA between the three connected elements. Thus, the detection and recognition of such non-genealogical bonds possibly yield deep consequences for evolutionary knowledge.

Evolutionary Thinking Beyond Genealogical Bonds. The systematic analysis of M-P3 patterns in networks suggests that one should assign comparable ontological importance to nonhomologous gene transfers in both single lineages and phylogenetically mosaic units to broaden the analysis of four types of evolutionary questions.

First, the origin of evolutionary novelties is generally considered through the impact of (selective/selected) mutations and recombination in nucleotide sequences within a genome (69) or random drift in populations. Although a number of mutations in key regulatory nodes might produce quite complex phenotypes, this focus must be expanded to solve the problem of how big novelties are acquired (e.g., how assembly of original combinations of preexisting, often unrelated biological entities increases diversity at every level of biological organization) (70, 71). A compelling example can be found in the recent expansion of a bacterial gene blaCTX-M-15, which inactivates most modern cephalosporin antibiotics in *Escherichia coli*. The ancestral gene of this detoxifying enzyme was a housekeeping gene in an organism ecologically accessible by *E. coli* and its plasmids, captured by an insertion sequence, and then moved into plasmids that were captured by particular cosmopolitan *E. coli* clones, including the widespread high-risk clone ST131-O25:H4-B2, which contributed to its worldwide spread. The *blaCTX-M-15* gene was then captured by new plasmids, which were captured in turn by other *E. coli* clones. Because some of these clones are particularly suited to be integrated in the intestinal microbiome of different types of animals, the *blaCTX-M-15* gene expanded multidimensionally, finally reaching even the hemolytic–uremic *E. coli* O104 responsible for food poisoning in Germany in 2011 (72–74). Consideration of M-P3s, the true binding of unlike to at least the origin of original evolutionary units, explicitly includes such evolutionary quantum leaps in studies of evolutionary novelties.

Second, introgressive and vertical descent can enrich models pertaining to the Darwinian threshold (75) (i.e., the time at which cellular lineages acquired sufficient autonomy, as lineages, to diverge from each other). After this threshold was crossed, the bonds of homology became more striking than the structures produced by M-P3s, but homology is not the only guideline to explain this early transition in the history of life. Considerations of vertical descent alone suggest that the more recent common ancestor of life would be more ancient than the Earth (76, 77), which seems impossible. Introgressive descent can, therefore, also contribute to understanding of early evolution. Interlevel mergers and multilevel clubs were likely key elements in the pre-Darwinian world (78). Investigations of ancient evolution should benefit from research to define the pool of shared genes of early multilineage and multilevel clubs rather than hinge on the definition of the single minimal cellular genome inferred from genealogical bonds between extant cellular beings. Unless introgressive descent is acknowledged, there will be Lost Common Ancestors: the contemporary mosaic evolutionary units of the hypothetical last common ancestor.

Third, the origin of lineages is often considered as a problem of branching order on a tree. However, genetic assortments crossing lineages and levels also yield lineages of major evolutionary players. The entry of eukaryotes on the scene, whether as the product of some sort of fusion (56) or successive endosymbioses...
multilineage units is possibly no less fundamental than the origin of multicellular life. Both phenomena require explaining how distantly related entities (e.g., cells or mobile elements) reach their current level of integration and the mechanisms deployed for passing on traits that belong to the complex rather than particular individuals or lineages. Similar questions can be raised for multilevel organizations. Thus, comparative analyses of multiple multilineage/multicellular clubs could identify convergent mechanisms, features, genomic properties, ecological affinities, or functional capacities of the members of such clubs. Analyses of M-P3s can set up an analytical framework to define the possible rules (85, 86) (the grammar of associations between the different entities), even in the absence of genealogical continuity.

Conclusions

Richard Owen proposed that instances of the same organ under every variety of form and function should be considered homologs. Darwin proposed a genealogical cause for that homology. He, thus, established a hidden bond particularly suited to diagnose and explain evolution of single lineages. Ever since that time, biologists have preferentially investigated evolutionary changes through relationships of homology and tree-like genealogical patterns. However, increasingly many evolutionary units and transitions seemed to depend on and arise from non-tree like processes. In particular, the analysis of the evolution of mergers and clubs requires us to uncover other bonds, reaching beyond strict kinship and beyond one biological level. Because in- trogressive descent structures biodiversity in ways that vertical descent does not, it seems essential to study the patterns caused by intersections and genetic exchanges between lineages (and not just within lineages). By starting with patterns as simple as M-P3s, it should be possible to improve our understanding of past, present, and future biological evolution significantly and encourage the inclusion of additional evolutionary units in our description of biological evolution. This line of research can expand the study of biological complexity beyond the usual genealogical bonds, revealing additional sources of biodiversity, and promote additional developments of the analytical apparatus required for network analysis to handle even more complex patterns generated by introgressive descent. We commend it to our readers.

ACKNOWLEDGMENTS.

The work of F. Bouchard is funded by the Social Sciences and Humanities Research Council of Canada. The research of F. Baquero is sponsored by the Seventh Framework Programme of the European Union (PAR-241476 and EvoTAR-282004) and the Carlos III Institute Research Fund (FIS-P10-02588). J.O.M. is funded by the Science Foundation Ireland Research Frontiers Programme (09/RFP/E02510).