Eukaryotic Pangenomes

Guy-Franck Richard

Eukaryotic Pangenomes

Guy-Franck Richard

Abstract The first eukaryotes emerged from their prokaryotic ancestors more than 1.5 billion years ago and rapidly spread over the planet, first in the ocean, later on as land animals, plants, and fungi. Taking advantage of an expanding genome complexity and flexibility, they invaded almost all known ecological niches, adapting their body plan, physiology, and metabolism to new environments. This increase in genome complexity came along with an increase in gene repertoire, mainly from molecular reassortment of existing protein domains, but sometimes from the capture of a piece of viral genome or of a transposon sequence. With increasing sequencing and computing powers, it has become possible to undertake deciphering eukaryotic genome contents to an unprecedented scale, collecting all genes belonging to a given species, aiming at compiling all essential and dispensable genes making eukaryotic life possible.

In this chapter, eukaryotic core- and pangenomes concepts will be described, as well as notions of closed or open genomes. Among all eukaryotes presently sequenced, ascomycetous yeasts are arguably the most well-described clade and the pangenome of *Saccharomyces cerevisiae*, *Candida glabrata*, *Candida albicans* as well as *Schizosaccharomyces* species will be reviewed. For scientific and economical reasons, many plant genomes have been sequenced too and the gene content of soybean, cabbage, poplar, thale cress, rice, maize, and barley will be outlined. Planktonic life forms, such as *Emiliana huxleyi*, a chromalveolate or *Micromonas pusilla*, a green alga, will be detailed and their pangenomes pictured. Mechanisms generating genetic diversity, such as interspecific hybridization, whole-genome duplications, segmental duplications, horizontal gene transfer, and single-gene duplication will be depicted and exemplified. Finally, computing approaches used to calculate core- and pangenome contents will be briefly described, as well as possible future directions in eukaryotic comparative genomics.

G.-F. Richard (✉)
Institut Pasteur, Department Genomes & Genetics, Paris, France

CNRS, Paris, France
e-mail: gfrichar@pasteur.fr

H. Tettelin, D. Medini (eds.), *The Pangenome*,
https://doi.org/10.1007/978-3-030-38281-0_12

Keywords Eukaryotes · Evolution · Fungi · Haptophyte · Mamiellale · Eudicotyledon · Monocotyledon · Protostomian · Metazoan

1 The Origin of Eukaryotes

Respiratory-competent eukaryotic cells emerged more than 1.5 billion years ago, from the endosymbiosis of an alphaproteobacterium and an ancestral archaebacterium, probably belonging to the Asgard clade (Zaremba-Niedzwiedzka et al. 2017). This protoeukaryote evolved, concomitantly, a complex system of membrane compartments that would ultimately lead to the isolation of the genomic content within a real nucleus (*eu karyon* in Greek) while the degenerated alphaproteobacteria gave rise to the mitochondria (López-García and Moreira 2006). The subsequent acquisition of photosynthesis through endosymbiosis with a cyanobacteria evolved this primitive cell into a protoalga from which all plants will eventually develop. The general outline of this scenario has been postulated for more than a century (Mereschowsky 1999; Sagan 1967) and modern-day DNA sequencing techniques allowed to precisely identify bacteria most closely related to modern eucaryotes, hence representing their most probable ancestors. However, the exact order of events is still a matter of debate among evolution specialists. Did membranes come first, to isolate nucleic acid metabolism from protein and sugar metabolism? Did the mitochondria come first, providing a considerable source of oxidative energy to further develop a complex network of membranes? These two scenarios are not necessarily exclusive and one may also imagine that a number of different protoeucaryotes emerged at roughly the same time (at geological scale) and competed with each other within similar ecological niches, until one lineage arose and was eventually selected to give rise to all eukaryotic life.

Given the bacterial origin of nucleated cells, it was assumed that most if not all eukaryotic gene families would share homology to prokaryotic genes. However, the sequencing of an old deep-branching eukaryote, the excavata *Naegleria gruberi* (Fig. 1), revealed that only 57% of its 4133 protein families had a clear prokaryotic homologue. The remaining genes showed no homology to bacterial sequences and therefore appear to be eukaryote inventions. Therefore, one must expect eukaryotic pangenomes to be significantly different from any known prokaryotic pangenome.

2 Sequencing Eukaryotic Genomes

Modern-day eukaryotes are estimated to represent 8,740,000 land species and 2,210,000 ocean species, for a total of roughly 11 million, one order of magnitude above procaryotes (Mora et al. 2011). Higher estimates, based on plankton sampling, suggest figures around 16 million of oceanic eukaryotes and 60 million of land species (de Vargas et al. 2015). Eukaryote classification is a complex problem taking

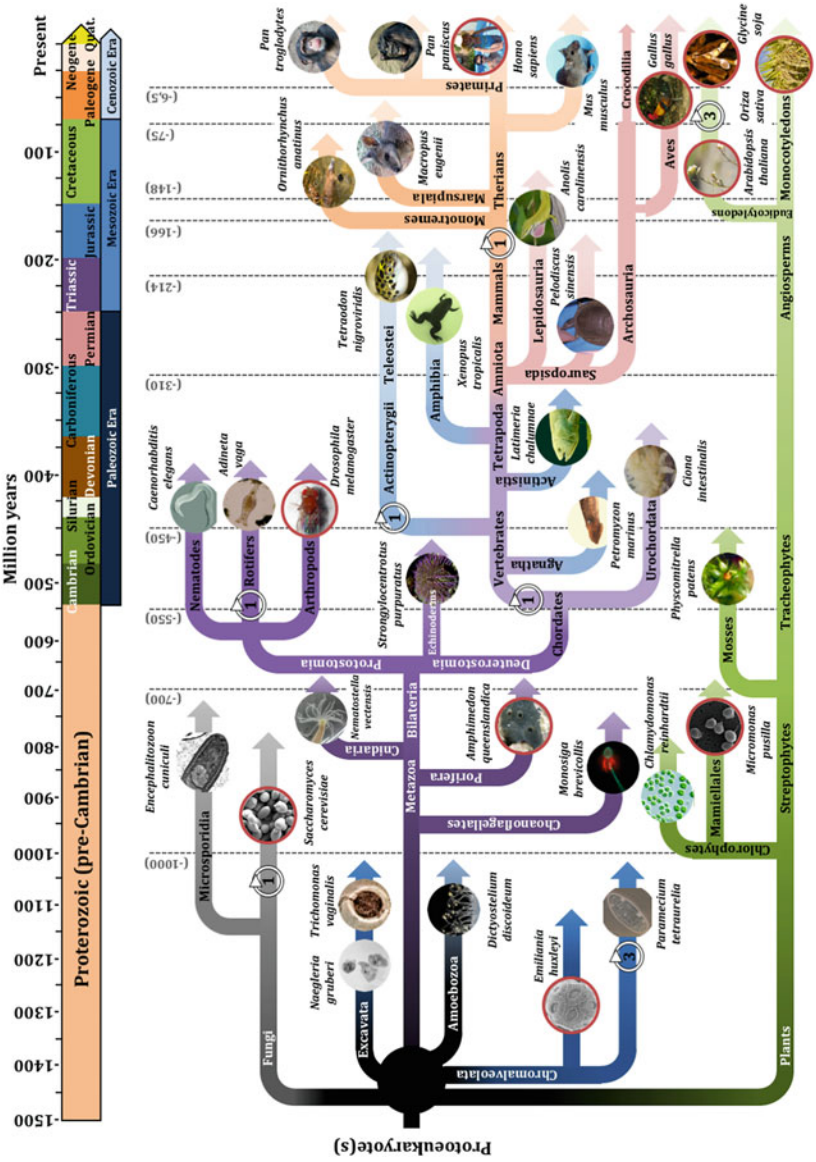

Fig. 1 Overview of some of the most representative eukaryotic genomes sequenced. At the top is shown a timeline in million years before present (note that the scale is different before and after Cambrian). Geological periods are indicated, as well as their corresponding eras. Each arrow represents one monophyletic

its roots into the nineteenth century zoology and botanics, but more recently gained much insight from whole-genome sequencing and molecular phylogeny reconstruction methods (Felsenstein 2004). Early eukaryotes (or old eukaryotes), such as fungi, monocellular green algae, excavata (one of the most basal lineage), amoebozoa, and chromalveolata diverged probably between 1.2 and 1.45 billion years ago (Embley and Martin 2006). Younger eukaryotes, like vertebrates, emerged 450 million years ago (Erwin et al. 2011), whereas *Homo sapiens* is still in evolutionary infancy with an estimated date of divergence from chimpanzee around 6.5 million years ago (Green et al. 2010) (Fig. 1).

The ascomycete *Saccharomyces cerevisiae* was the first eukaryote whose nuclear genome was totally sequenced, more than 20 years ago (Goffeau et al. 1996). In the 1990s, it took the efforts of 633 scientists from more than 100 laboratories during 8 years to complete it (Goffeau et al. 1997). In the modern genomic era, sequencing is fast, cheap, and allows to decipher whole eukaryotic genomes at unprecedented scale and pace in human history. At the present time, 707 different eukaryote species, including 54 unicellular animals (Protozoa) or algae, 300 metazoans (multicellular animals), 137 plants, and 216 fungi had their genome sequenced to various levels of completion and assembly. Indeed, the actual pace at which eukaryotes are being sequenced is so elevated, that the aforementioned figures will be completely outdated when this book will be published. Remarkably, one of the most ambitious current genome projects envisions to sequence all eukaryotic life present on planet Earth, and the cost of such a project would be similar to what was spent to sequence the first human genome alone (Pennisi 2017). Some of the most representative eukaryote species, whose genomes were completely sequenced are represented in Fig. 1, on the evolutionary branch they belong to, along with their estimated geological period of appearance based on molecular clocks.

Fig. 1 (continued) group (or clade) that survived to present day. Branch lengths are arbitrary. When more than one organism was sequenced in a given clade, only one was shown (for example, among all sequenced bird genomes only the paradigmatic *Gallus gallus* species was represented). Vertical dotted lines indicate speciation time from the most recent common ancestor, calculated from molecular clocks. For example, Actinopterygians (bony fish) separated from other vertebrates approximately 450 million years ago. Note that Precambrian radiation datations were only tentatively attributed, given the large uncertainties associated to ancient eukaryotes. Circled numbers represent whole-genome duplications detected by sequencing. The constriction between Archosaurians and Aves represents the Archaeopteryx, the ancestor of all modern birds (Hillier et al. 2004). The smaller arrow between Archosauria and Crocodilia represents the dinosaurian mass extinction, 66 million years ago, among whom the only survivors were the ancestors of modern-day crocodiles (Brugger et al. 2017; Renne et al. 2015). Red circled species were used to define core- and pangenomes and are more extensively described in the text

3 The 1000 Genome Projects

One of the most remarkable aspects of modern-day genomics is the ambition to describe a large number of individuals (usually in the range of thousands) belonging to the same monophyletic group (or clade). When the first eukaryotic genome sequences were completed, it became apparent that one genome would not be sufficient to describe the whole species. Several programs subsequently started, aiming at sequencing a large number of individuals belonging to the same species and comparing them to the first genome, usually called "reference genome" because its state of completion and annotation was often more advanced. Several of these projects have been completed over the last few years: 1011 *S. cerevisiae* genomes (Peter et al. 2018), 1135 *Arabidopsis thaliana* genomes (The 1001 Genomes Consortium 2016), 2504 followed by 10,545 human genomes (Telenti et al. 2016; The 1000 Genomes Project Consortium 2015), and 1483 rice genomes (Yao et al. 2015) have already been sequenced, but complete analyses of gene content and core- and pangenome calculations are not always published. Even more ambitious endeavors are planned: the 10,000 plant genome project led by the Chinese BGI[1] aims at sequencing one representative plant from every major clade (Normile et al. 2017); the same institute launched in 2015 the 10,000 bird genome project, in an attempt to sequence every one of the 10,500 living bird species (Zhang 2015). The i5K initiative is planning to sequence 5000 arthropod genomes (i5K Consortium 2013) or the Genome 10K project intends to sequence 10,000 vertebrate genomes (Genome 10K Community of Scientists 2009). All these projects—and many others to come—will contribute to unraveling the complete set of genes used by eukaryotic life forms on Earth. With this wealth of data at hand, assuming it will not be too overwhelming for available data storage and computing power, essential questions should find their answers. What are the core genes shared by all eukaryotic species? How many different versions of the same gene (alleles) can be found? How many variable or dispensable genes can be detected in a given species? What is the size of a species pangenome, of a clade pangenome, of the eukaryotic pangenome itself?

4 Defining Eukaryotic Pangenomes: Open or Closed?

The very notion of pangenome was coined by Hervé Tettelin and colleagues in a 2005 seminal article, describing sequencing and genome analysis of eight strains of *Streptococcus agalactiae*. Despite a high degree of synteny[2] between isolates, the authors detected 69 genomic islands that were absent in at least one genome, some characterized by an atypical nucleotide compositional bias, suggestive of a possible acquisition by horizontal transfer. They showed that the number of shared genes in all

[1] Beijing Genomics Institute, the largest—by far—sequencing center in the world.

[2] Synteny: gene order along a chromosome.

species decreased at each addition of a new genome, reaching the minimal number of 1806 genes. On the contrary, each genome addition increased the number of variable genes, those that are absent in one or more strain. They proposed that a bacterial species may be defined by a set of genes present in all strains (core-genome) and by a dispensable—or variable—set of genes, composed of those present in at least one strain but absent from all others. The addition of these variable genes to the core-genome would make what was called the "pangenome" (from the Greek word *pan (παν)*, meaning "whole") (Tettelin et al. 2005). Mathematical modeling showed that the pangenome measurement followed the Heap's law, an empirical law used in information retrieval, in which as more and more books are read, the number of different words grows as a power law of the total number of books read. The function form of the power law depends on two parameters: the exponent α and a proportionality constant. Practically, the number of new genes discovered after each new genome sequence will be: $n = \kappa N^{-\alpha}$, in which κ is a constant, N is the number of genomes sequenced, and $\alpha > 0$. For $\alpha > 1$, the pangenome size approaches a plateau as more and more genomes are sequenced, the pangenome is "closed" (Fig. 2a). On the other hand, for $0 < \alpha \leq 1$, the pangenome size will increase at each new genome addition and the pangenome is "open" (Fig. 2b) (Tettelin et al. 2008).

Among sequenced bacterial species, some exhibit a closed pangenome, for example *Staphylococcus aureus* ($\alpha = 1.84$), *Streptococcus pyogenes* ($\alpha = 1.88$), *Ureaplasma urealyticum* ($\alpha = 2.5$) or the extreme case of *Bacillus anthracis* ($\alpha = 5.6$). Others display an open pangenome, like *Bacillus cereus* ($\alpha = 0.65$) or the cyanobacteria *Prochlorococcus marinus* ($\alpha = 0.80$). Note than when α is equal or very close to 1, the pangenome is still open, but the rate of acquisition of new genes is very slow. This is the case of *Escherichia coli* ($\alpha = 1.04$), *Streptococcus agalactiae* ($\alpha = 1.05$), or *Streptococcus pneumoniae* ($\alpha = 0.98$) (Tettelin et al. 2008).

5 Yeast Pangenomes

5.1 Saccharomyces cerevisiae

Historically, budding yeast was the first eukaryote whose genome was completely sequenced (Goffeau et al. 1996). A British collaborative work in which 70 *S. cerevisiae* and *S. paradoxus* isolates were sequenced to low coverage showed that *S. cerevisiae* strains showed less variability than *S. paradoxus* strains. Worldwide budding yeast population structure was made of a few geographically isolated lineages and of several mosaic genomes, and underlined the possibility that humans played a major role in producing these variations by transporting and selecting yeast strains (Liti et al. 2009). Following this pioneering work, a collaborative effort of two French laboratories and the Genoscope led to the completion of 1011 *S. cerevisiae* isolates, collected worldwide, from domesticated, wild, or human origin (mainly clinical). This sequencing effort allowed to determine that Chinese and Taiwanese

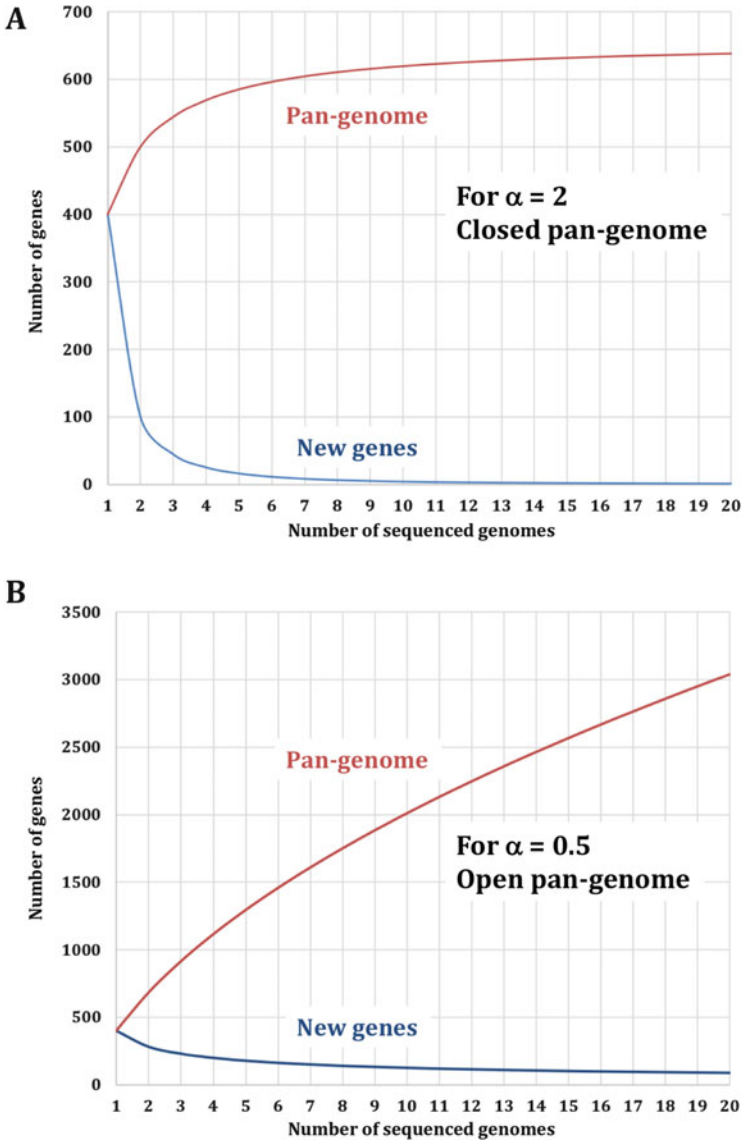

Fig. 2 Open versus closed pangenomes. (**a**) Closed pangenome. In this example, the number of new genes = 400× (Nbr genomes)$^{-\alpha}$, with $\alpha = 2$. The number of new genes revealed by each new genome sequence rapidly decreases and the pangenome size reaches a plateau. (**b**) Open pangenome. The number of new genes = 400× (Nbr genomes)$^{-\alpha}$, with $\alpha = 0.5$. The number of new genes revealed by each new genome sequence keeps on growing and the pangenome size steadily increases

strains were closer to *Saccharomyces paradoxus* and to the root of the *Saccharomyces sensu stricto* than strains from any other origin, strongly supporting a single out-of-China origin for *S. cerevisiae*, that subsequently spread all over the planet. Using de novo assembly and a specific detection pipeline, it could be determined that the yeast core-genome contained 4940 Open Reading Frames (ORFs) whereas 2856 ORFs were variable within the population, for a total of 7796 ORFs constituting the pangenome (Peter et al. 2018) (Table 1). Core ORFs were mostly found in one copy per haploid genome, while ca. 20% of variable ORFs were absent or present in more than one copy. The authors subsequently looked at the origin of these variable ORFs and classified them in three different groups, based on their phylogeny: ORFs with their closest ortholog in another *S. cerevisiae* strain and consistent with genome phylogeny were considered as being ancestral acquisitions; ORFs with their best ortholog in another *Saccharomyces* species were considered to be introgressions; and finally ORFs more related to another yeast species outside the *Saccharomyces* complex were treated as horizontal gene transfers (HGT) (Fig. 3a). Using these definitions, 1380 variable ORFs were assigned to an ancestral inheritance, 913 were designated as introgressions, and 183 were likely to be the result of HGT events from distant relative yeast species. Half of these HGT ORFs could be traced to *Torulaspora* or *Zygosaccharomyces* species. Given that these yeasts share similar environmental fermentative niches, it is likely that such physical promiscuity favored frequent transfer of genetic material between these species. In six cases, large HGT events (38–165 kb) were identified, but most isolates retained only mosaics of small segments suggesting that the large ancestral HGT underwent several rounds of successive deletions leading to the complex patterns observed today. Among the 913 introgressions, 97% were unambiguously acquired from *S. paradoxus*, all *S. cerevisiae* ORF carrying at least one *S. paradoxus* ORF, suggesting continuous gene flows between these two yeast species. This is in good agreement with a former work using microarrays to genotype *Saccharomyces* strains of different origins, in which most introgressions detected in *S. cerevisiae* came from *S. paradoxus* (Dunn et al. 2012). Finally, two-thirds of ancestral acquisitions were present in at least half the yeast isolates, suggesting that they segregated in most strains since the time of their acquisition (Fig. 3b).

The core- and pangenomes of the S288C reference strain were analyzed more thoroughly for variable gene functions. Out of 6081 ORFs, 1144 were identified as variable. The distribution of these ORFs was found to be skewed toward subtelomeric regions, which have been known for a long time to be highly polymorphic among yeast strains and species (Fabre et al. 2005). Functions of variable ORFs were strongly enriched for cell-wall and membrane components, cell–cell interactions, and secondary metabolism. Finally, core-genome ORFs were found to exhibit lower levels of loss-of-function mutations, as compared to pangenome ORFs, as well as a lower dN/dS ratio of nonsynonymous over synonymous substitutions, showing that the former were less constrained than the latter.

Table 1 Core- and pangenome contents

Clade	Species (or genus)	Isolates	Core-genome	Variable genes[a]	Pangenome	Status[b]
Saccharomycotina	*Saccharomyces cerevisiae*	1011	4940	2856 (37%)	7796	ND
	Candida glabrata	33	3603	9915 (73%)	13,000–14,000	ND
	Candida albicans	21	6069	120 (2%)[c]	6189	ND
Taphrinomycotina	*Schizosaccharomyces*[d]	4	4218	782 (16%)	5000	ND
Eudicotyledon	*Glycine soja*	7	28,716	30,364 (61%)	50,080	Open
	Brassica oleracea	9	49,895	11,484 (19%)	61,379	ND
	Populus trichocarpa	6	≈34,000	12,000–13,000 (26%)	46,000–47,000	Closed
	Arabidopsis thaliana	19[e]	26,373	11,416 (30%)	37,789	Open
Monocotyledon	*Oriza sativa*	66	26,372	16,208 (38%)	42,580	Closed
	Zea mays	503	16,393	25,510 (61%)	41,903	Closed
	Hordeum vulgare	16	10,922	17,840 (62%)	28,762	Closed
Mamiellales	Three different species[f]	4	7137	2824 (23%)	12,518	ND
Haptophyte	*Emiliania Huxleyi*	14	20,055	10,514 (34%)	30,569	ND
Protostomian	*Drosophila*[g]	12	6698	40,852 (86%)	47,550	ND
Metazoans	*Homo sapiens*	5[h]	ND	ND	ND	Open

ND: Not Determined by the authors and not possible to calculate from published data

[a]The proportion of variable genes as compared to the pangenome size is indicated in parenthesis

[b]Open or closed pangenome (see Fig. 2 and text)

[c]Calculated from an average value. The real number of variable genes might be slightly larger

[d]*S. octosporus*, *S. pombe*, *S. japonicus* and *S. cryophilus*

[e]More than 1100 *A. thaliana* genomes were sequenced, but 19 transcriptomes were used to determine core-and pangenome contents (see text)

[f]Two isolates of *Micromonas pusilla*, plus *Ostreococcus tauri* and *Ostreococcus lucimarinus*

[g]The 12 sequenced genomes corresponded to 12 *Drosophila* species, not 12 isolates from the same species

[h]More than 10,000 human genomes are available, but 5 of them serve as references (see text)

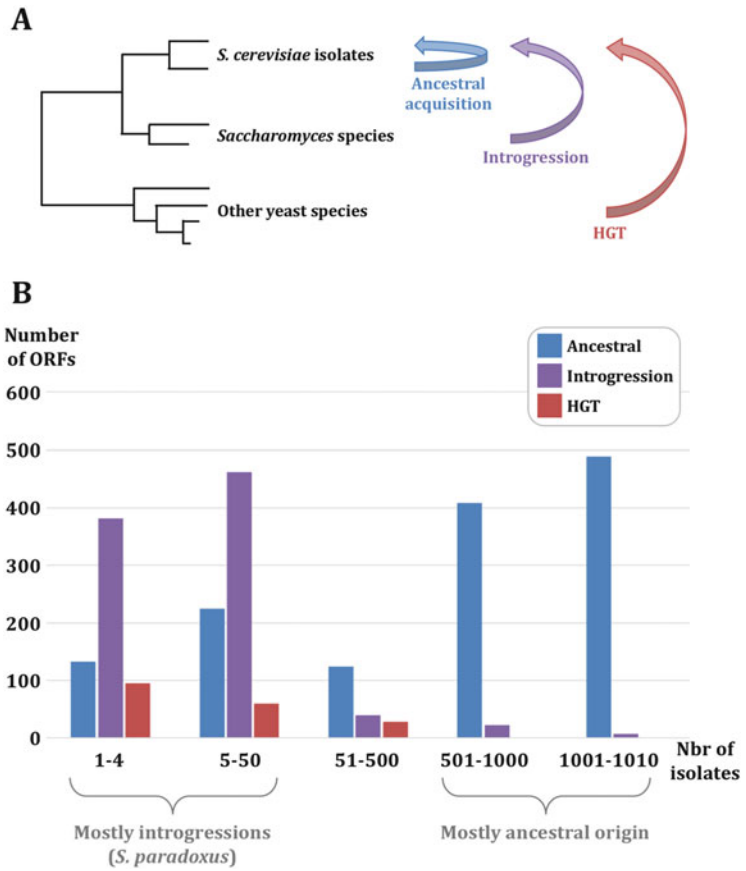

Fig. 3 Variable ORFs of the *S. cerevisiae* pangenome. (**a**) Phylogenetic origin of variable ORFs. ORFs were considered ancestral acquisitions when the best match was found to be a *S. cerevisiae* ORF (blue arrow), it was treated as an introgression when the best homolog was another *Saccharomyces* species (purple arrow), or a horizontal gene transfer (HGT, red arrow) when it was found to be another yeast species. (**b**) Distribution of variable ORFs. The number of isolates is indicated on the *X*-axis and the number of variable ORFs in each category is represented on the *Y*-axis

5.2 Candida glabrata

C. glabrata is an opportunistic pathogen responsible for candidiasis and bloodstream infections in immunocompromised patients (Bodey et al. 2002). It is the second cause of nocosomial infections, after *Candida albicans*, and a growing concern in public health, due to its resistance to azole antifungal drugs (Pfaller and Diekema 2004). Despite its genus name, its genome is closer to *S. cerevisiae* than to *C. albicans*. It belongs to the *Nakaseomyces* clade that also includes *Candida nivariensis* and

Candida bracarensis, two emerging pathogens, as well as *Nakaseomyces delphensis*, *Nakaseomyces bacilisporus*, and *Candida castellii*, three nonpathogenic species (Fig. 4). Comparison of orthologous proteins conservation shows that this clade is as distant from the *Saccharomyces* clade as man is distant from fish (Dujon 2006). Hence, the distance between orthologous proteins belonging to these two monophyletic groups is similar to the distance covered by vertebrate proteins since the actinopterygian radiation, some 450 million years ago[3] (Fig. 1). *C. glabrata* exhibits frequent chromosome polymorphisms among different isolates, due to translocations, copy number variations (CNV), gene tandem amplifications (Muller et al. 2009), formation of neo-chromosomes (Polakova et al. 2009), and the presence of many large tandem repeats known as megasatellites (Rolland et al. 2010; Thierry et al. 2008, 2009). The five aforementioned pathogenic and nonpathogenic *Nakaseomyces* species were sequenced to high coverage and their sequence was compared to the *C. glabrata* CBS138 reference strain (Dujon et al. 2004). Protein contents range from 4875 for *C. castellii* to 5315 for *C. bracarensis*, figures significantly lower than the 5886 *S. cerevisiae* proteins (Gabaldon et al. 2013). Among gene losses in *Nakaseomyces*, four entire multigene families (*PHO*, *SNZ*, *SNO*, and *PAU*) were absent in all species or represented by only one member in *C. castellii* or *N. bacillisporus*. These genes are involved in phosphate metabolism (PHO), in nutrient limitation response (SNZ and SNO), or in alcoholic fermentation (PAU). The loss of BNA genes, functioning in de novo synthesis of nicotinic acid probably results from the yeast adaptation to its human host, since colonization of the urinary tract occurs through induction of adhesin genes, upregulated in nicotinic acid-poor medium, such as urine (Domergue et al. 2005). The *C. glabrata* genome contains a large number of genes that are absent from *S. cerevisiae* and specifically involved in adhesion and virulence. The *EPA* genes, a family of glycosyl-phosphatidylinositol cell-wall genes, completely absent from *S. cerevisiae*, was represented by 18 members in the *C. glabrata* reference strain (CBS138), and seven additional genes were present in the BG2 strain, widely used in adhesion studies (Cormack et al. 1999). Remarkably, the two other pathogenic species, *C. bracarensis* and *C. nivariensis*, contained respectively 12 and 9 members of the EPA family, whereas the nonpathogenic *N. delphensis* and *C. castelli* harbored respectively one and three copies and *N. bacillisporus* presented only one distant homologue. In addition, the *C. glabrata* genome contained 44 genes comprising internal repeats, whose motifs were 135–300 nt long, tandemly repeated 3–30 times in frame (Thierry et al. 2008). These megasatellites encode many serine and threonine residues and genes harboring these tandem repeats were proposed to encode cell-wall glycoproteins and to be involved in cellular adhesion (Thierry et al. 2009). Phylogenetic studies of 21 fungal genomes showed that these megasatellites were uniquely found in *C. glabrata*, but their presence among other members of the *Nakaseomyces* has not been tested yet (Tekaia et al. 2013).

[3]This does not mean that *Saccharomyces* and *Nakaseomyces* diverged 450 million years ago, because there is no reliable molecular clock for yeasts.

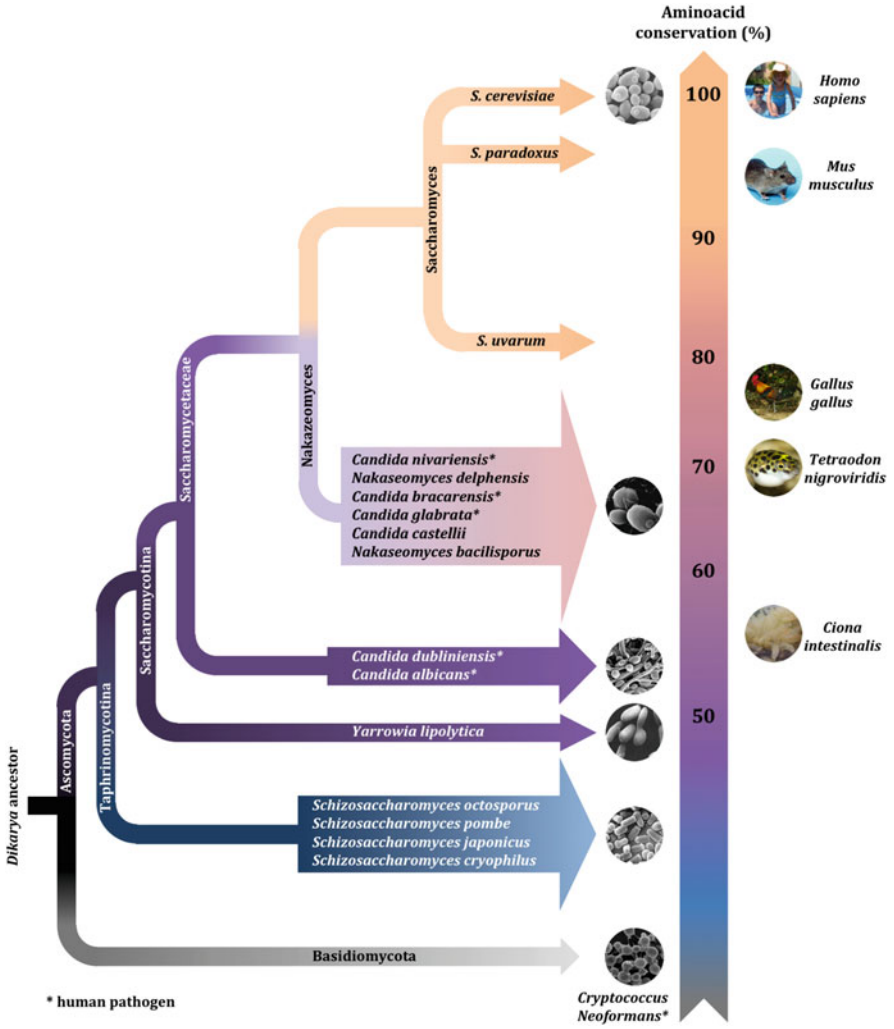

Fig. 4 Yeast pangenomes of the *Dikarya* tree. On the left, the figure shows some of the yeast species whose genomes were completely sequenced, arranged by clade. Branch lengths are arbitrary and do not reflect evolutive distances. On the right, amino acid conservation of orthologous proteins between yeast and between animal species are indicated (adapted from Dujon 2006)

In a very recent study, 33 isolates of *C. glabrata* of different geographical origins were fully sequenced and compared to the CBS138 reference strain (Carreté et al. 2018). Altogether, 108 genes were deleted or duplicated in these strains, half of them encoding glycosylphosphatidylinositol-anchored adhesin homologues, showing the extensive variability of this gene family within this clade. The core-genome contained 3603 proteins, significantly less than for *S. cerevisiae* (see above). On the contrary, the number of variable ORFs was higher than budding yeast, since

302–580 predicted genes (mean: 342) were found to be unique of each isolate, for a total of 9915 strain-specific genes among 29 strains considered.[4] This figure may be partially overestimated, due to automated annotations or clustering artifacts, but from these data one may infer that the *C. glabrata* pangenome covers 13,000–14,000 genes, almost twice as many as the *S. cerevisiae* pangenome.

In conclusion, yeasts of the *C. glabrata* clade contain significantly fewer genes than *S. cerevisiae*, with specific gains and losses as compared to their distant cousin. However, gene content is highly variable among *Nakaseomyces* and the *C. glabrata* pangenome size is larger than the *S. cerevisiae* pangenome, although further analyses are needed to narrow down these numbers.

5.3 Schizosaccharomyces *Genomes*

Fission yeasts are very distant relatives of *S. cerevisiae* and the *Taphrinomycotina* clade comprise only four known species: *Schizosaccharomyces japonicus*, *Schizosaccharomyces cryophilus*, *Schizosaccharomyces octosporus*, and the model yeast *Schizosaccharomyces pombe*. They form a basal branch of the *Dikarya*[5] tree (Fig. 4) and exhibit very distinct life history and metabolism as compared to *Saccharomycotina*. Under many aspects, *S. pombe* is actually closer to metazoans than to budding yeasts: among the more prominent features, large repetitive centromeres, heterochromatin histone methylation, heterochromatin proteins, RNA interference, telomere-binding proteins, cell-cycle control, the mitochondrial translation code, splicing and spliceosome components are more similar to metazoans. In addition, core orthologous genes in *S. pombe* are closer to metazoan genes than to other *Ascomycota*. Phylogeny reconstruction of the clade using high coverage sequence of the four *Schizosaccharomyces* species and 440 single-copy core orthologues surprisingly revealed that *S. pombe* and *S. japonicus* were as far to each other (55% average amino acid identity) as man and *Ciona intestinalis*, an urochordate (Fig. 1) (Rhind et al. 2011). The two other species, *S. octosporus* and *S. cryophilus*, were closer to each other (85% amino acid identity). Retrotransposons are numerous in *S. japonicus* and sequence divergence of their reverse transcriptase suggests that they predate the last ancestor of the *Ascomycota*. However, transposons were dramatically lost in the three other species, since *S. pombe* harbors two related retrotransposons, *S. cryophilus* contains only one and *S. octosporus* only has sequence relics of reverse transcriptase sequences. This loss was accompanied by a reorganization of centromere architecture, replacing the numerous transposons found at *S. japonicus* centromeres by other kinds of repeated sequences unrelated to transposons and specific of each of the other three species.

[4]Four isolates were excluded from this analysis because of low-quality assembly.

[5]*Ascomycota* and *Basidiomycota* together form the *Dikarya*.

Out of ≈5000 coding genes in fission yeasts, 4218 (84%) were identified as single-copy orthologues common to all four species. For some gene families, the level of conservation was even higher: 93% of protein kinases were common and more surprisingly 81% of introns (2901 out of 3601) were identical across the clade. Most gene gains were species- or clade-specific genes not found in another yeast species, whereas gene loss included the glyoxylate cycle, glycogen biosynthesis, the phosphoenolpyruvate carboxykinase, fewer *ADH* genes and lack of transcriptional regulators of glucose repression, all these changes reflecting the inability of fission yeast to use ethanol as a carbon source, although it produces it by fermentation. Hence, despite large evolutionary distances of conserved orthologous proteins, *Schizosaccharomyces* show a remarkably stable gene content, supporting a pangenome size only 10–20% larger than its core-genome.

5.4 Candida albicans

Candida albicans is another opportunistic pathogen, responsible for mucosal and systemic infections in immunocompromised patients. It is also a commensal of the gastrointestinal tract. Natural isolates of *C. albicans* are diploid and under specific conditions they are able to mate, resulting in tetraploid cells subsequently shifting to diploidy via random chromosome loss (Bennett and Johnson 2003). The nuclear genome of SC5314, a standard laboratory strain widely used in molecular analyses, was published in 2004. It revealed a high level of single-nucleotide polymorphisms (SNP) between both homologues, representing 90% of all detected polymorphisms, with an average frequency of one SNP in 237 bases. Heterozygosity was not homogeneous, since several chromosomes were interrupted by large regions of homozygosity (Jones et al. 2004). After that initial study, 21 clinical isolates of *C. albicans*, characterized by different phenotypic profiles, were also completely sequenced. Single-nucleotide polymorphisms were very limited among the isolates, being one order of magnitude lower than what was commonly found among *C. glabrata* strains (Gabaldón and Fairhead 2019). The gene content of these isolates was very similar to that of SC5314 reference strain, since most of its genes were present in all isolates (6069 genes out of 6189—or 98%—on the average), with few variable genes (Table 1). Genes exhibiting the most variable number of copies were retroelements as well as the subtelomeric *TLO* gene family. The position and number of *TLO* genes varied from 10 to 15 among isolates, indicative of a high level of plasticity (Hirakawa et al. 2015). More recently, the *Candida dubliniensis* genome, another opportunistic pathogen, less virulent than *C. albicans*, was sequenced. Except for translocations and chromosomal rearrangements that may be expected between two yeast species, both gene contents were found to be surprisingly similar. Out of 5569 orthologues, 5363 (96.3%) were more than 80% identical at the nucleotide level, and synteny was conserved for 98% of genes (Jackson et al. 2009). The search for species-specific genes identified 111 ORFs in *C. dubliniensis* and 191 in *C. albicans*. However, most of these variable ORFs corresponded to transposable

elements. When these were filtered off, the real number of species-specific genes dropped to 29 and 168, respectively. Among those, the *TLO* gene family (12 members in *C. albicans*) was specifically expanded in this species, since only two copies were detected in *C. dubliniensis* and species-specific copies were monophyletic, supporting an independent expansion in *C. albicans*. On the contrary, the *IFA* gene family (13 members in *C. albicans*) underwent massive gene loss in *C. dubliniensis*, since several gene relics at various stages of decay were identified in this yeast species. In conclusion, in the present state of analysis, it appears that the core-genome common to *C. albicans* and *C. dubliniensis* probably approximates 5400 genes and that their pangenome may be predicted to be slightly larger, possibly around 6200 genes.

6 Plant Pangenomes

6.1 Soybean Genomes

Glycine max is the cultivated soybean variety, whose genome was published in 2010 (Schmutz et al. 2010). It was domesticated 5500 years ago and has been under intensive selection by human populations for yield increase. It diverged from the wild variety, *Glycine soja*, 800,000 years ago, well before its domestication. Therefore, natural selection contributed to differentiation of the two subspecies well before human selection started. In order to estimate the genetic diversity between domesticated and wild soybean species, the genome of seven *Glycine soja* isolates from south-east Asia were sequenced and compared to each other and to *G. max* (Li et al. 2014). Gene number ranged from 54,256 to 57,631, depending on the isolate and hundreds of genes were identified as gained or lost as compared to domesticated soybean. The *G. soja* core-genome contained 28,716 genes, while 30,364 variable genes were identified. Most of them (58%) were shared by two to six out of seven samples, whereas 12,916 (42%) were uniquely found in one of the seven isolates. The pangenome therefore contained 50,080 genes and covered 986.3 Mb of sequence. Its size increased with each new isolate, but it did not reach an asymptote, suggesting that adding new isolates would increase pangenome size (Fig. 2). Interestingly, dispensable genes exhibited more sequence variability than core genes. SNP frequency was at 2.67 sites per kilobase for variable genes, whereas it was significantly higher for core genes (4.12 sites per kilobase), and a similar bias was found for indels. Biological processes enriched in dispensable genes include specific metabolic processes, antioxidant activity, and structural molecule activity. These genes were also less conserved than core genes since 58% could not be assigned to a functional annotation, as compared to only 34% of the core genes. Lineage-specific genes include 11 genes implicated in effector-triggered immunity, acting as pathogen detectors, reflecting adaptation to various biotic stresses.

The domesticated soybean genome contains 1794 genes involved in acyl lipid metabolism, illustrating the effect of its intense selection for oil and fatty acid

production. Among those, 32 exhibited CNV when compared to *Glycine soja*, 252 contained SNPs or indels and 21 showed high dN/dS ratios, suggestive of their possible positive selection in *Glycine max*.

In conclusion, *G. soja* pangenome was found to be twice as large as its core-genome, and its comparison with the domesticated *G. max* species revealed the effect of human selection on this widely cultivated crop.

6.2 Rice Genomes

Rice (*Oryza sativa* L.) is one of the most important crops in the world, feeding half the world population. The genome sequence of this monocotyledon was published in 2005 (International Rice Genome Sequencing Project 2005), although draft sequences of each chromosome were released earlier. Domesticated rice comprises two subspecies: *indica* and *japonica*. The reference genome (Nipponbare) is a *japonica* subspecies and contains 37,544 protein-coding genes, among which 2859 (8%) seemed to be uniquely found in rice. In an effort to explore the genetic diversity of cultivated rice, 1483 sequences of both subspecies from 73 countries, sequenced at low coverage (1–3 X), were compared to the reference genome. Comparison of both subspecies sequences to the reference genome identified 8991 predicted genes for the dispensable *indica* genome and 6366 for the *japonica* genome. Among these, strong evidence of expression or high homology was found for 1120 genes of the *japonica* dispensable genome and 1913 genes of the *indica* dispensable genome. Out of these 1913 high confidence genes, 1189 (62%) contained a recognizable protein domain, for a total of 276 different protein domains altogether (Yao et al. 2015).

In a more recent study, 66 isolates of cultivated rice as well as wild rice (*Oriza rufipogon*)[6] were sequenced to high coverage and the corresponding genomes were de novo assembled and compared (Zhao et al. 2018). Chromosomal introgressions from *indica* were detected in ≈16% of tropical *japonica* genomes. Numerous insertions and deletions were identified within genes, since a total of 10,872 genes were at least partially absent from the reference genome, due to large indels. Protein-coding genes present in at least one isolate were annotated and all transposable elements were filtered out. A total of 26,372 genes were found to be common to more than 60 rice isolates and were therefore considered to constitute the rice core-genome. Variable genes, present in less than 60 genomes, were assigned to a dispensable set of 16,208 genes, so that the rice pangenome reached a total of 42,580 genes. A larger proportion of core proteins (78%) than of dispensable proteins (36%) matched to known domains, suggesting that some of these variable genes may be pseudogenes or artifacts. Among dispensable genes, abiotic and biotic response genes, controlling disease resistance in rice were found to be enriched. When coding genes were sequentially added from each genome, the number of

[6] 28 *Oriza sativa japonica*, 25 *Oriza sativa indica*, and 13 *Oriza rufipogon* isolates.

different genes reached a plateau, although more pronounced for gene families than for singletons. This strongly suggests that the rice pangenome is almost closed and that further sequencing of rice isolates will not prove to be very useful in identifying new dispensable genes (Table 1).

6.3 Maize Genomes

Transcriptome sequencing of polyadenylated mRNAs was used in a genome-wide study as a proxy to determine the complete set of protein-coding genes within 503 diverse maize inbred isolates of different origins (Hirsch et al. 2014). RNA-seq reads were mapped to the *Zea mays* reference genome and reads that did not match were used for identification of novel transcripts. To limit redundancy, only the longest transcript of each locus was taken into consideration for further analysis. A total of 8681 high confidence transcripts that were absent from the reference genome were categorized as dispensable genes. Among those, 50% matched with rice and sorghum proteins, ruling out that they could be artifacts or contaminants. Transcripts detected in all isolates, including the reference line, represented 16,393 genes and constituted the core-genome. Dispensable transcripts, that were identified in only a subset of isolates, represented 25,510 genes, for a pangenome of 41,903 genes, very close to the rice pangenome, although the proportion of variable genes was much higher in maize (61% vs. 38% for rice). Sequential addition of genes belonging to each isolate revealed that the number of different singletons and gene families reach a plateau (more pronounced for singletons), demonstrating that the maize pangenome was closed, or very close to completion (Table 1).

6.4 Cabbage Genomes

Brassica oleracea is a diploid eudicotyledon, comprising remarkably morphologically diverse crops, including cabbage, cauliflower, broccoli, Brussels sprout, kohlrabi, and kale. The *B. oleracea* pangenome was built by sequencing nine isolates (eight cultivated and one wild—*Brassica macrocarpa*) and anchoring them on one of the two reference genomes (Parkin et al. 2014). The assembled pangenome covers 587 Mb and represents 61,379 genes, after removal of transposable elements. The core-genome constitutes the majority of the pangenome, representing 49,895 genes (81%), whereas 11,484 genes (19%) are variable, 1322 (2%) being present in only one line. Dispensable genes were enriched for functions predicted to be involved in disease resistance, defense response, water homeostasis, amino acid phosphorylation, and signal transduction. Lineage-specific variable genes comprised biotic and abiotic stress response genes, similar to what was observed in rice and soybean. *B. oleracea* underwent a whole-genome triplication specific to this lineage, in which gene families involved in auxin function and in morphological variations were

amplified, these last ones perhaps contributing to the wide morphology diversity observed in this species.

There are 14 variable genes predicted to regulate flowering time and maturity in *B. oleracea*, but all of them were absent from one of the two reference strains (TO1000), a rapid cycler. One of the flowering loci, *FLC* (Flowering Locus C), is an important regulator of vernalization and regulates flowering time variation by the number of gene copies. One *FLC* gene was present in *Arabidopsis thaliana*, whereas four paralogues were found in *B. oleracea*. All four were part of the core-genome and two additional homologues were detected: one was present in all lines except the TO1000 reference strain and the other was present only in *B. macrocarpa* and one isolate (Cauliflower1). Independent functional studies showed that disruption of this gene in cauliflower led to early flowering, strongly suggesting that its absence in TO1000 was responsible for the early flowering of this rapid cycler (Golicz et al. 2016).

Genetic signatures of the core-genome and of the variable genome are very different. Core genes are longer on the average and harbor more exons. They also have lower mean SNP density and the ratio of non-synonymous over synonymous substitutions was lower than for variable genes, suggesting that core genes were under a more selective purifying selection than variable genes. In conclusion, *B. oleracea* core and variable genes exhibit the same properties that were observed in other eucaryotic pangenomes.

6.5 Poplar Genomes

The genome of *Populus trichocarpa*, black cottonwood, was published in 2006. Out of its predicted 45,555 protein-coding genes, 40,307 (88%) had a homologue in *Arabidopsis thaliana*, while conversely 91% of *A. thaliana* predicted genes showed some similarity to a *P. trichocarpa* gene (Tuskan et al. 2006). More recently, six isolates of other poplar species, four *Populus negra* and two *Populus deltoides*, were sequenced to 26-45X coverage and compared to the *P. trichocarpa* reference genome. Genome comparisons identified 7889 deletions and 10,586 insertions in the two newly sequenced species, as compared to *P. trichocarpa*. However, a large majority of these were due to transposons and retrotransposable elements (62% of deletions and 84% of insertions), a feature shared by all plant pangenomes sequenced so far. Once transposon sequences were filtered out, 3230 genes exhibiting CNV signatures between at least two of the samples were detected. These CNVs were significantly more abundant within 3 Mb from telomeres and corresponded to gene additions or deletions in one or more sample. A total of 230 variable genes were detected among *P. nigra* samples, and of 174 dispensable genes between the two *P. deltoides* isolates. The reference *P. trichocarpa* genome showed 187 genic variations with *P. nigra* and 213 with *P. deltoides*. Among these dispensable genes, 70% belonged to a gene family, allowing to detect some over-represented gene functions. Remarkably, variable genes were preferentially involved

in signal transduction, receptor activity, and disease resistance, similarly to what was observed for soybean, rice, and cabbage (Pinosio et al. 2016).

The authors of this study calculated that the poplar pangenome was approximately 500 Mb, 80% being shared by all the isolates and therefore constituting the core-genome. When *P. nigra* and *P. deltoides* genomes were compared to the reference *P. trichocarpa*, 2270 genes were absent from at least one sample and 2453 other genes were detected in a variable number of copies, for a total of 4723 variable genes. Unfortunately, the proportion of dispensable genes between *P. nigra* and *P. deltoides* was not determined, and it was therefore not possible to figure out the exact size of the poplar pangenome. However, estimates suggest a size of ≈34,000 genes for the core-genome and ≈12,000–13,000 variable genes, giving a pangenome size of ≈46,000–47,000 genes. Using available data about *P. nigra* dispensable genes, it is tempting to suggest that its pangenome should be closed.

6.6 Mamiellales Genomes

Micromonas pusilla is a marine picoeukaryote of the Mamiellales order, measuring less than 2 μm and living in all oceans worldwide. Two independent isolates of *M. pusilla* were sequenced and their genomes were compared to those of *Ostreococcus lucimarinus* and *Ostreococcus tauri*, two other Mamiellales. Surprisingly, the two *Micromonas* shared only 90% of their 10,000 predicted genes, whereas the two *Ostreococcus* shared 97% of theirs. Comparison of the four sequences allowed to define a core-genome containing 7137 genes, involved in photosynthesis, hydroxyproline-rich glycoproteins (essential components of plant cell-wall), and meiosis genes. These were unexpected since Mamiellales are generally considered to be asexual, suggesting that these genes were remnants of their common ancestor with land plants, or alternatively that they possessed a kind of sexuality that has not been described yet. This last hypothesis would be compatible with the presence of glycoproteins known to be expressed after sexual fusion in *Chlamydomonas reinhardtii*. In addition to core genes, 14% of proteins (1384) were shared by both *Micromonas* isolates but were not found in Ostreococcus. These include enzymes for plastid peptydoglycan synthesis. These "shared" genes were found to evolve more rapidly than core genes. A large proportion of genes present in only one of the two *Micromonas* isolates exhibited homology to animal or bacterial lineages, supporting their acquisition by horizontal transfer. Altogether, 793 and 826 genes were unique to each of the two *Micromonas* isolates, 689 were specific of *O. tauri* and 249 were unique to *O. lucimarinus*. These variable genes when added to the 7137 core genes and to the 2824 genes shared by at least two of the four genomes, gave a Mamiellales pangenome size of 12,518 genes (Nordberg et al. 2014; Worden et al. 2009).

7 Animal Pangenomes

7.1 Drosophila *Genomes*

Drosophila melanogaster is one of the most intensively studied animal models. The first draft of its genome was published in 2000 (Adams et al. 2000). Its euchromatin part covered ≈120 Mb and contained 13,600 genes, only twice as many as budding yeasts. Following this pioneering work, 11 other fly species originating from Africa, Asia, the Americas, and the Pacific islands were sequenced and compared to *D. melanogaster* reference genome. Gene numbers range from 13,733 for *D. melanogaster*[7] to 17,325 for *Drosophila persimilis*. Sequence comparisons established that 49% of *D. melanogaster* genes were conserved as single-copy orthologues across the whole set of species, defining a set of 6698 core genes. Collectively, the 12 *Drosophila* genomes contain 40,852 variable genes, for a pangenome size of 47,550 genes, but unfortunately it was not possible to determine if this pangenome was closed or open with published data. However, some interesting observations were made. First, effector proteins (like antimicrobial peptides) evolved by rapid duplications and deletions and were significantly underrepresented in the core-genome. Second, gene families forming most of the variable gene content expanded or contracted at a rate of one fixed gene gain or loss every 60,000 years. Common functions among some of the rapidly evolving families include defense response and proteolysis. Third, the vast majority (98%) of *Drosophila* proteins were ancestrally present at the root of the genus. Out of the 296 non-ancestral proteins, 252 were specific of the *Sophophora* subgenus or were complex acquisitions. The remaining 44 genes were lineage-specific (four of them are found only in *D. melanogaster*), were shorter than the average, harbored fewer introns and 40% of them (18/44) were testis-specific, consistent with previous observations about new *Drosophila* genes (Drosophila 12 Genomes Consortium 2007).

In conclusion, *Drosophila* core-genes represent roughly 40–50% of each species gene pool and variable genes arise most of the time by duplication or deletion of an existing gene, with very little de novo gene creation.

7.2 *Avian Genomes*

Birds encompass the richest variety of species among tetrapod vertebrates, with more than 10,000 different species. In an international effort, 48 avian species, covering most avian clades were sequenced to low or high coverage and compared to the existing three reference genomes (zebra finch, turkey, and chicken), as well as to three crocodilian genomes, the closest bird relatives. After filtering for transposable elements, each genome was predicted to contain ≈14,000–17,000 genes. They

[7]Gene number was refined since publication of the draft genome sequence.

contained a low level of repeated elements (4–10%) as compared to other tetrapods (34–52% in mammals, for example).

Genes responsible for morphological and physiological peculiarities of the clade were analyzed more in depth. Flight capacity was permitted through duplication and positive selection of genes regulating skeleton morphology and bone development. Out of 89 genes involved in ossification half of them showed traces of positive selection, compared to one-third of the 31 orthologous genes in mammals.

Feathers are made of α- and β-keratins, the latter only found in birds and reptiles. Aves genomes contained fewer α-keratin genes as compared to mammals but the repertoire of β-keratins has expanded (up to ≈150 copies in zebra finch). Similarly, most avian genomes contained a higher number of opsin genes than mammalian genomes, partly explaining their more advanced visual system. Genomic elements that were highly conserved among the 48 bird genomes were identified genome wide. Such elements covered 11 Mb (1% of the avian genome) and were significantly underrepresented in coding regions. Actually, the proportion of conserved elements in noncoding regions were 50-fold higher and mostly corresponded to regulatory regions of developmental genes. This result suggested that few avian-specific genes arose in this clade, most of the genomic changes resulting from differences in developmental regulations (Seki et al. 2017).

In conclusion, avian genomes are smaller than mammalian genomes, both in size and in number of genes, due to extensive deletions of chromosomal segments in the ancestral lineage. More precise analyses are now required to sort out core genes from dispensable ones in order to be able to define core- and pangenome sizes and contents.

7.3 Human Genomes

The first human genome drafts were published in 2001 at the same time by the Human Genome Sequencing Consortium and by Celera Genomics (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), and a more complete version of the academic sequence was released in 2004 (International Human Genome Sequencing Consortium 2004). A few years later, James Watson's own genome was deciphered (Wheeler et al. 2008), rapidly followed by the first Asian genome (Wang et al. 2008) and the first African genome (McKernan et al. 2009). The human pangenome was built from comparisons between the NCBI human reference genome and four genomes: Venter's (Celera Genomics), Watson's, YH (Asian genome), and NA18507 (African genome), as well as individual human sequences retrieved from GenBank. Four types of sequence variants were detected: (1) sequences that were frequent in African populations but rapidly declined out of Africa; (2) sequences that were rare in African populations but became more frequent with geographical distance; (3) sequences that were present at a low frequency in European populations; and (4) sequences that were rare in Asian populations. This analysis led to the conclusion that the human pangenome should

include 19–40 Mb of additional sequence in addition to the reference genome and that complete coverage of all gene variants should be achieved with the sequencing of 100–150 randomly sampled individuals, worldwide. Analysis of sequences that could not map to the reference genome showed that some of the most abundant genes were those encoding *DUX* homeobox proteins (113 hits in YH and 58 in NA18507), known to be associated with chromatin. Also very frequent were gene families known to be rapidly evolving, such as mucins, zinc-finger proteins, and olfactory receptor proteins (Li et al. 2010).

In conclusion, the present-day human pangenome is still open and will require many more finished sequences in order to be resolved. No doubt that recent efforts to sequence 1070 Japanese genomes (Nagasaki et al. 2015), 2504 individuals from 26 worldwide origins (The 1000 Genomes Project Consortium 2015) or 10,545 human genomes representative of the main human populations (Telenti et al. 2016) should allow to more precisely define human core- and pangenomes and definitely solve this question.

7.4 Reaching for the Metazoan Pangenome

With a wealth of more than 300 metazoan genomes sequenced, defining a core- and a pangenome for multicellular animals could seem a reachable goal. However, with an estimation time for the last common ancestor of all metazoans around 800 million years ago (Erwin et al. 2011), identification of a reliable set of core genes might prove challenging. The sponge *Amphimedon queenslandica* is an early metazoan (Fig. 1) whose genome was sequenced in 2010. It is predicted to contain 18,693 protein-coding genes. Comparison with 4670 metazoan gene families defined a set of 1286 proteins that seem to be metazoan specific, thus defining a draft core-genome for multicellular animals (Srivastava et al. 2010). Many gene expansions observed in the metazoan lineage arose by subsequent tandem or local gene duplications, but extensive work is now needed in order to extract this information from available metazoan genome sequences.

8 The Oceanic Pangenome

The TARA ocean program aims at sampling all planktonic lifeforms of the world's ocean (de Vargas et al. 2015). Metatranscriptomes were established from high-coverage polyA RNA-Seq performed on 441 size-fractionated planktonic communities. Subsequent clustering created a nonredundant set of 116 million transcribed sequences, at least 150 bases long. Despite the sampling effort, it was calculated that 166–190 million sequences would be needed to reach saturation of all oceanic eukaryotic expressed sequences. Half of these sequences had no match in public databases, suggesting that they may correspond to new genes, but most of these (60%) were present as single copies. Transcription of these new genes showed that

they were expressed to the same level as known families, suggesting that they were conserved in a smaller number of species or that they were present in less abundant taxonomic groups. Increasing the sampling effort should solve this issue (Carradec et al. 2018). These data, although preliminary and not totally exhaustive, demonstrated that it was possible to extract thousands of new eukaryotic genes belonging to yet uncharacterized species from large oceanic metagenomes. It would be difficult to use the same approach for land eukaryotes for which a comprehensive sampling will be much more tedious and time consuming.

8.1 The Haptophyte Alga Emiliania huxleyi

Marine phytoplankton is responsible for carbon fixation and export to the sea floor as calcite, as well as carbon dioxide release during the calcification process. Their influence on carbon metabolism and export to the deep ocean is complex and crucial for the Earth ecosystem. The haptophyte *E. huxleyi* CCMP1516 reference genome was determined, as well as 13 other isolate genomes, from subarctic to tropical oceanic origins (Read et al. 2013). Repetitive elements were extremely abundant, representing about two-thirds of the sequence and include retrotransposons (1%), DNA transposons (3%), rDNA-related repeats (3%), paralogous genes (10%), tandem repeats and low complexity regions, especially 10–11 bp tandemly repeated minisatellites (34%) and unclassified repeats (16%). These repetitive elements account for a large part of the considerable genome size variability, that ranges from 99 to 133 Mb between isolates (141.7 Mb for the CCMP1516 reference). The reference genome gene content was then compared to three isolates of very distant origins.[8] Out of 30,569 predicted genes in the reference, a total of 5218 (17%) were absent from at least one of three isolates and 364 were missing from all three. Further comparisons with the other isolates strengthened this conclusion: the core-genome contained 20,055 genes, about two-thirds of the reference genes, whereas the remaining genes were variable, making *E. huxleyi* pangenome a complex gene repertoire. Besides repeated elements, the genome encodes many iron-binding proteins, 80 in the core-genome and 30 as variable genes. Iron is essential for calcification and photosynthesis and these differences probably reflect ecological disparities among isolates. In addition, the *E. huxleyi* pangenome encodes 700 proteins whose function relies on metal binding: selenium (49 proteins, 20 gene families), zinc (413 proteins), or copper (65 proteins). Finally, the pangenome contains 26 genes involved in vitamin metabolism, but is unable to synthesize vitamins B_1 and B_{12}, restricting *E. huxleyi* to oceanic regions where these are freely available. In conclusion, the large pangenome of this haptophyte is probably necessary to accommodate its ubiquitous distribution in oceans and illustrates physiological and morphological disparities observed among isolates.

[8]English channel, north-eastern pacific ocean and Great Barrier reef.

9 Where Do Eukaryotic Variable Genes Come From?

At the present time, there are six independent origins for novel eukaryotic genes: interspecific hybridizations, whole-genome duplications, segmental duplications, horizontal gene transfer, single gene duplication, and de novo gene creation (Fig. 5).

9.1 Interspecific Hybridizations

The American botanist Edgar Shannon Anderson published in 1949 a book describing interspecific hybridizations between flowering plants and genotype combinations resulting from these crosses (Anderson 1949). Since then, it became widely accepted among botanists that such events were frequent among plants, resulting in frequent transfers of genes from one species to another. Interspecific hybridizations were very common among yeast species too (Morales and Dujon 2012). Modern brewing yeast, *Saccharomyces pastorianus*, is the offspring of two successive hybridizations, an ancestral one between *Saccharomyces uvarum* and an unknown species and a more recent one between the resulting hybrid and *S. cerevisiae* (Nguyen et al. 2011).

Despite these interesting observations, zoologists were stuck with a very conservative notion of species, based on reproductive isolation, i.e., two species were considered as different if the offspring of their mating was sterile. This remarkably conservative thinking did not take into consideration that many natural fertile interspecific animal hybrids were already described: liger (lion and tiger), pizzlies (polar bear and brown bear), Hawaiian duck (mallard/Laysan duck), *Heliconius* butterflies (*Heliconius cydno* and *H. melpomene*) and Darwin's finches, to name only a few (Pennisi 2016). However, this very conservative way of thinking hit a wall when genome-wide sequencing of ancient human DNA demonstrated that modern *Homo sapiens* were the result of at least two interspecific hybridizations. The first one occurred 50,000–80,000 years ago between *Homo neanderthalensis* and ancestral *Homo sapiens*, after their "out of Africa" journey. This resulted in the retention of 1–4% of Neanderthal genes in all modern *Homo sapiens* genomes, except for those of pure African descent (Green et al. 2010). The second hybridization occurred between offsprings of *Homo sapiens* and *Homo neanderthalensis* and a new species of ancestral human, the Denisovan man (named from the cave in the Siberian Altai mountains in which it was discovered). The hallmark of this hybridization can still be seen in present-day Melanesian populations in which 4–6% of genes come from this ancestral Denisovan man (Prüfer et al. 2014). More recently, the same team discovered the remnants of a 13-year-old girl who was the daughter of a Neanderthal mother and of a Denisovan father, demonstrating that these two ancient human populations also hybridized with each other, around 50,000 years ago (Slon et al. 2018).

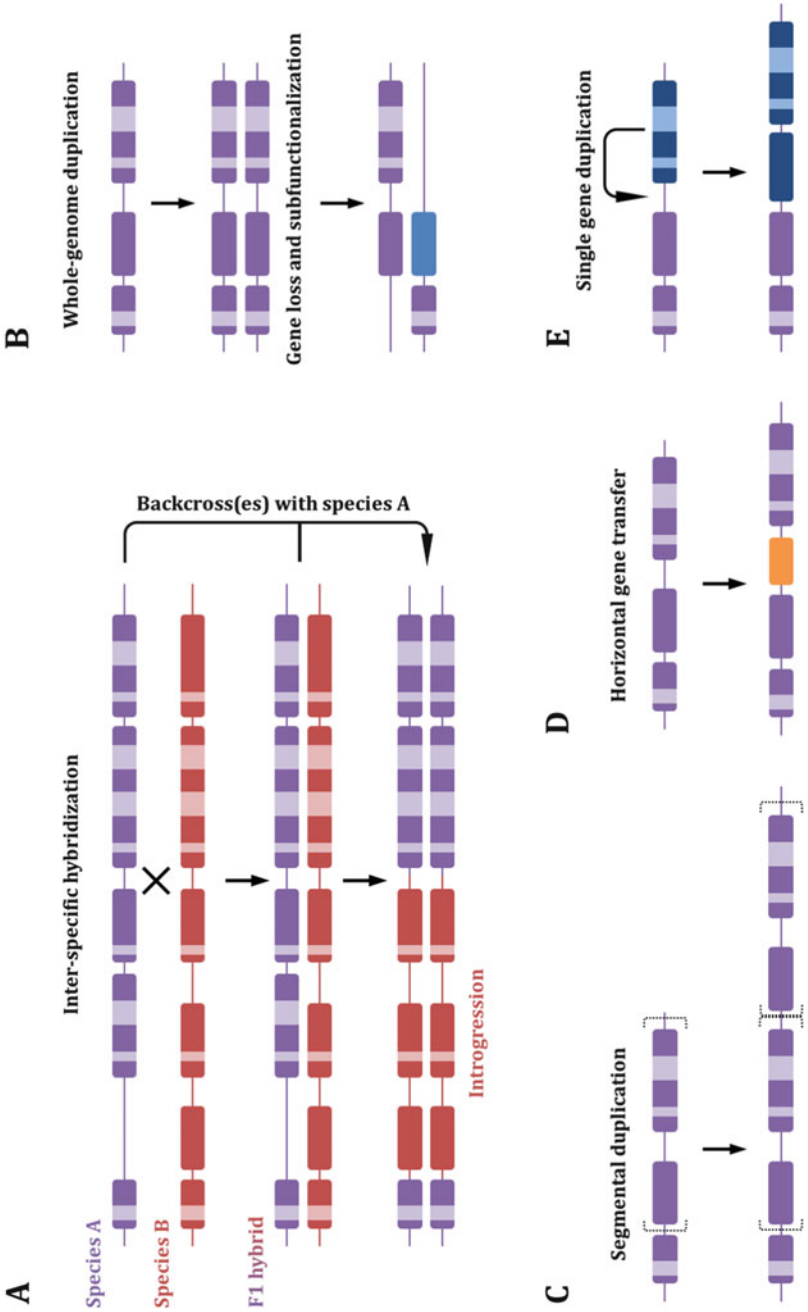

Fig. 5 Gene innovations in eukaryotes. Color boxes represent exons, lighter boxes are introns. (**a**) Interspecific hybridization. Two germ cells with different genomes will merge and produce a fertile hybrid. Several rounds of backcrossing with one of the parents (Species A) will homogenize the genotype but may end

Successive hybridizations can be detected as chromosomal introgressions, large DNA fragments which may be fixed by natural selection following backcrossing between an hybrid and one of its parents (Fig. 5a). One such example in modern humans comes from the Tibetan population. Their genome contains a transcription factor induced under hypoxic conditions, *EPAS1*, whose expression correlates with hemoglobin levels in low atmospheric oxygen pressure. This gene is located in a 120 kb chromosomal region containing a large number of SNPs that were very common in Tibetan and Denisovan DNA, but found at very low frequencies in Han Chinese genomes. This proved that adaptation to high altitude in Tibetan populations was due to a large chromosomal introgression inherited from their Denisovan ancestry (Huerta-Sánchez et al. 2014).

At the present time, it is safe to admit that interspecific hybridizations have been a significant source of gene novelty in eukaryotic genomes, from fungi to animals and plants. However, if living species may mate with other species living in a close ecological niche and produce a fertile offspring, we should now define species independently of the outdated reproductive barrier. Indeed, one may ask what is a species?

9.2 Whole-Genome Duplications

Compared to interspecific hybridizations, bringing together two distinct sets of genes, whole-genome duplications bring together two exact same sets of genes (Fig. 5b). Whole-genome duplications were extremely frequent in every branch of the eukaryote tree, in ascomycetes (Dujon et al. 2004; Kellis et al. 2004; Wolfe and Shields 1997), in paramecium (Aury et al. 2006), in teleostean fish (Jaillon et al. 2004), plants (International Wheat Genome Sequencing Consortium 2014; Jaillon et al. 2007; Vision et al. 2000), rotifers (Flot et al. 2013), and vertebrates (Dehal and Boore 2005), just to cite a few (Fig. 1). These whole-genome duplications were rapidly followed by extensive gene loss, in order to restore gene dosage, but some of the duplicated genes—also called onhologues—may be maintained for a longer time and

Fig. 5 (continued) up in selecting a chromosomal region from the other parent (Species B) that will become a permanent introgression. It is possible that other mechanisms besides backcrossing may generate chromosomal introgressions. (**b**) Whole-genome duplication will be followed by extensive gene loss to counteract gene dosage defects. Sub- or neofunctionalization may occur on one of the two onhologues. Only one chromosome was represented for the sake of clarity, but all chromosomes are duplicated in this process. (**c**) Segmental duplication of a large chromosomal segment (in brackets) may produce several duplicated genes in a single event. (**d**) A gene (in orange) may be transferred from another organism. Horizontal gene transfer may also affect a small number of genes. (**e**) A gene is reversed transcribed and the cDNA integrated in the genome. Former introns are possibly lost in the process if reverse transcription occurs on a spliced transcript. Note that an allelic transposition is represented but ectopic duplications are frequent

evolved new functions by neo- or subfunctionalization. *S. cerevisiae* harbors two copies of cytochrome c resulting from an ancestral whole-genome duplication, one encoded by the *CYC1* gene and the other by *CYC7*. The latter is expressed when oxygen levels are so low that cells are in hypoxia, whereas the former is expressed when oxygen levels are normal, a classic case of subfunctionalization (Downie et al. 1977). An interesting example of neofunctionalization was discovered in an Antarctic fish, the eelpout *Lycodichthys dearborni*, whose genome contains two SAS genes, resulting from an ancient duplication. Both SAS-A and SAS-B genes encode an enzyme involved in sialic acid biosynthesis. SAS-B got subsequently partially duplicated and the resulting paralogue was deleted for four out of six exons, making a much shorter gene. The resulting protein happened to bind more efficiently ice crystals than the full-length protein, interfering with crystal growth and behaving as a good antifreeze protein. Subsequent tandem amplifications of this shorter version of SAS-B gave the eelpout the ability to resist extreme cold conditions (Deng et al. 2010).

It might prove technically difficult to discriminate between a recent whole-genome duplication and an interspecific hybridization between two closely related species, without a good reference. It is possible that some chromosomal duplications that were thought to arise from whole-genome duplications were actually acquired by hybridization. In a near future, the achievement of more and more eukaryotic genomes originating from the same clade should eventually dismiss any concern about the origin of close paralogues.

9.3 Segmental Duplications

Another frequent source of novelty comes from local or ectopic duplication of a chromosomal DNA segment, called segmental duplication (Fig. 5c). Their length range from a few to several hundreds of kilobases and they have been found in every eukaryotic species sequenced so far. They are also commonly called copy-number variations (or copy-number variant, or CNV) since their copy number may vary from one genome to another, or structural variant (SV). Spontaneous segmental duplications were found in the yeast *S. cerevisiae*, during experimental evolution of a wild-type strain (Dunham et al. 2002) or using a gene dosage assay for growth recovery (Koszul et al. 2004). These chromosomal duplications could be sometimes quite large, covering 41–655 kb. It was subsequently demonstrated that the mechanism generating segmental duplications was break-induced replication (BIR), a replication-based recombination process that could involve homologous sequences or microhomologies at the junction of duplicated segments (Payen et al. 2008).

Segmental duplications were also described in mouse (Bailey et al. 2004), in primate genomes (Cheng et al. 2005), as well as in man (Bailey et al. 2002). They are known to be associated with several human disorders (Emanuel and Shaikh 2001) and most of them were found to have recently emerged in human history (Jiang et al. 2007). They are undoubtedly a source of gene novelty by successive duplications of

large chromosomal segments, although their impact on gene content diversity has not been precisely evaluated yet.

9.4 Horizontal Gene Transfer

Very common between prokaryotes, horizontal gene transfer of a gene (or of a small number of genes) was limited to a few examples in eukaryotes, but may be more widely spread than previously thought. Such events have been identified among *Saccharomycetaceae* yeasts (Fig. 4). Out of 255 species-specific genes, 11 were identified as possible gene transfers from bacterial species, based on sequence similarities and reconstructed phylogenies (Rolland et al. 2009). In *S. pombe*, 34 genes were identified as good candidates for horizontal transfer from bacteria, 16 having occurred before radiation of the clade, 9 being specific to *S. pombe* (Rhind et al. 2011).

Sexuality is a natural obstacle to the propagation of a horizontally acquired gene to metazoan offspring since it must become integrated in the germ line. Nonetheless, some remarkable examples of gene transfer between bacteria or yeast to animal genomes have been described. *Wolbachia pipientis* is a symbiotic bacteria living inside several arthropods and some nematodes. Its genome sequence led to the discovery that 44 out of 45 *Wolbachia* genes were indeed integrated in the genome of the tropical fruit fly *Drosophila ananassae*, one of the natural hosts of this bacteria. Among the other species subsequently screened for the presence of *Wolbachia* genes, one nematode, one mosquito, one tick, three wasps, and five *Drosophila* species contained DNA fragments of various lengths originating from the bacteria (Dunning Hotopp et al. 2007).

Another striking example is the horizontal transfer of yeast genes to pea aphid (*Acyrthosiphon pisum*). This insect displays a red-green color polymorphism that serves to escape its natural predators. The different colors are due to different forms of carotenoid pigments found in individuals. Animals require carotenoids for several essential functions but they are unable to make them. Therefore, they normally find them in their diet. Remarkably, seven carotenoid synthases and carotenoid desaturases, enzymes required for pigment biosynthesis, are encoded by the aphid genome. Comparisons with existing sequences showed that these genes cluster with orthologues from fungi species and subsequent experiments led to the conclusion that these genes were transferred from a fungal pathogen or aphid symbiont, at the root of the aphid clade, followed by subsequent duplications of the transferred gene (s) (Moran and Jarvik 2010).

One last example comes from bdelloid rotifers, near-microscopic animals found in freshwater habitats worldwide. They lost sexual reproduction due to a specific chromosomal organization incompatible with meiotic recombination (Flot et al. 2013). Telomeric regions of *Adineta vaga*, a bdelloid rotifer whose complete genome has been sequenced, revealed dozen of genes of foreign origin. These were found in large telomeric chromosomal segments covering tens of thousands

of nucleotides and encoding various proteins playing a role in sugar or amino acid metabolism, in intracellular oxydo-reduction, or in the synthesis of antibiotics and toxins. Most of these genes came from bacteria or fungi species, some of them may have been transferred from plants. Among genes that were identified as of bacterial origin, some harbored introns, whereas their bacterial counterpart did not, suggesting that introns were acquired after transfer from bacteria. Telomeric regions being also enriched in transposable elements, the role of transposons in these massive gene transfers is still an open question (Gladyshev et al. 2008).

9.5 Single-Gene Duplication

Single-gene duplications may occur as allelic or ectopic genome insertions. When occurring in allelic position, they led to tandem repeats of paralogous genes, and were found in variable numbers in eukaryotic genomes. In ascomycetous yeasts, a few dozen tandem gene arrays were detected in each species, mostly composed of two to three copies. However, the *Debaryomyces hansenii* genome contained no less than 247 arrays of tandem paralogues, distributed all over its genome, some of them counting eight or nine tandemly repeated copies (Dujon et al. 2004). Ectopic paralogous gene duplications were also very frequent events in eukaryotes. Most carry the hallmark of retrotransposition: lack of introns, presence of a 3′-end polyA tract and remnants of target site duplications. These retrogenes were also called retroposons (Brosius 1991) and the transposition mechanism was studied in *S. cerevisiae* (Schacherer et al. 2004) as well as in human cells (Esnault et al. 2000). It relies on the reverse transcription of a mature mRNA by a reverse transcriptase (encoded by L1 elements in human cells), followed by integration of the cDNA at an ectopic or allelic locus (Fig. 5e). These duplicated genes lack promoter sequences that were absent from the mature transcript and are therefore pseudogenes, unless they luckily transpose near an active promoter. The human genome contains approximately 10,000 retrogenes, including more than 1700 ribosomal pseudogenes, while the mouse genome contains more than 200 copies of glyceraldehyde-3-phosphate dehydrogenase and *Caenorhabditis elegans* genome harbors more than 2000 pseudogenes (reviewed in Richard et al. 2008).

Extensive retroposition was also frequently detected in plants, the rice genome containing 1235 retrogenes. Interestingly, only 337 (27%) were identified as pseudogenes containing premature stop codons or frameshifts. Subsequent experiments concluded that more than half of the remaining retroposons were probably functional genes. In addition, 380 out of 898 intact retrogenes harbor a chimeric structure containing a flanking exonic sequence (Wang et al. 2006). Therefore, contrarily to the human genome in which most retroposons are pseudogenes, retroposition in the rice genome seems to be an active process rapidly creating new functional genes.

9.6 De Novo Gene Creation

Some remarkable cases of de novo gene invention have been well documented, although the total number of such cases having occurred during evolution of eukaryotes is probably underestimated. *Alu* retrotransposons are very common in primate genomes, being found in more than 1,000,000 copies, covering ≈13% of the genome size and present in almost every protein-coding gene intron (International Human Genome Sequencing Consortium 2001). In dozens of reported cases, an *Alu* sequence was found to be spliced with an upstream exon, resulting in a chimeric peptide (Makałowski et al. 1994). These hybrid proteins are a source of genetic novelty, although their total number in the human genome has not been precisely determined yet.

Before eukaryotes, the living world was asexual, except for bacterial conjugation that may be considered as a very primitive form of mating. Differentiation between two sexes appeared with the first eukaryotic cells and was found almost universally in the eukaryotic world, suggesting that it must be an ancestral acquisition. Sexual reproduction starts with the fusion between two haploid gametes of opposite sex, one male and one female, called syngamy, followed by the merging of both genetical contents. It was recently discovered that the protein responsible for syngamy (called HAP2) was structurally and functionally related to a viral membrane fusion protein. HAP2 was conserved in plants and animals and must have been transferred from a virus to a common ancestor at the root of the eukaryotic lineage (Fédry et al. 2017).

Therian mammals include marsupials and placental (or eutherians), like mouse or man (Fig. 1). In eutherians, egg development takes entire place within the uterus and the placenta is larger and more elaborated than in marsupials. In humans, two genes were responsible for placenta growth, *syncitin-1* and *syncitin-2*. These genes both derived from an envelope protein gene captured from an ancestral virus 25–40 million years ago. Remarkably, the mouse genome harbored two homologues, *syncitin-A* and *syncitin-B*, also deriving from a viral infection in the murine lineage around 20 million years ago, but they are not orthologous to their human counterparts, showing that the placenta was independently invented twice in two mammalian lineages by a similar mechanism of viral gene capture (Dupressoir et al. 2009).

In *D. melanogaster*, the *Sdic* gene coding for a sperm-specific dynein chain was the result of a local duplication and a complex rearrangement between two genes: *Cdic* and *AnnX*. The resulting *Sdic* gene was transcribed from a neo-promoter located in an intronic sequence and the first 21 amino-acids of the resulting protein came from this same intron, now spliced as the first exon of the *Sdic* mRNA (Nurminsky et al. 1998).

One may argue that the above examples are not real de novo gene creations, since they rely on preexisting DNA sequences (*Alu* elements, viral genes, or serendipitous rearrangements of existing exons). It is remarkable that the genome of the excavata *Naegleria gruberi* (Fig. 1) contained 40% of genes without any obvious similarity to any bacterial gene, suggesting that they could be real de novo eukaryotic inventions

(Fritz-Laylin et al. 2010). However, it is possible that many genes that appeared to be novel have indeed diverged so much from their prokaryotic ancestor that they cannot be identified anymore. Hence, the hunt for real de novo gene creation promises to be exciting but seriously challenging!

10 Bioinformatics Tools to Calculate Core- and Pangenomes

Most pangenome analyses were so far performed on prokaryotic genomes. Computing tools rely on the initial determination of genes belonging to the core-genome, followed by addition of all variable genes to build the species pangenome. The initial step is crucial, since one wants to identify the exhaustive list of orthologues belonging to each of the species isolates. Orthologue identification generally uses bidirectional best hits (BDBH), or BLAST followed by a clustering algorithm such as MCL, or comparison of protein domains using Hidden Markov Models (HMM) (reviewed in Guimarães et al. 2015). In a slightly different approach, PanOCT used synteny information in addition to orthology to define the core-genome. The program used a "conserved gene neighborhood" information to discriminate real orthologues from very recently duplicated paralogues whose sequences are indistinguishable (Fouts et al. 2012).

Calculation of eukaryotic core- and pangenomes is significantly more complex for several reasons: (1) the abundance of transposable elements, including novel undescribed transposons absent from dedicated databases; (2) the morcellated nature of genes, particularly in young eukaryotes; (3) the presence of large gene families that make orthologue identification tedious; and (4) the relative incompleteness of genomic sequences, particularly of those containing numerous repeats. In an original approach trying to tackle these problems, genomic and transcriptomic data from 19 *A. thaliana* isolates were analyzed using the GET_HOMOLOGUES-EST software, designed to use tissue-specific expression patterns to build core- and pangenomes. Results support a set of 26,373 core genes and of 11,416 variable genes, for a pangenome containing a total of 37,789 genes. The pangenome is open, each new isolate adding approximately 70 novel variable genes. Core genes exhibit a higher expression level than variable genes and they are under stronger selective pressure ($dN/dS << 1$), confirming what was already observed in other eukaryotes. The same software was used to analyze transcriptomic data from 16 *Hordeum vulgare* isolates (barley), a monocotyledon plant. The barley genome is 34 times larger than *A. thaliana* (4 Gb vs. 119 Mb) and contains 75% of repetitive elements. Its core-genome contains 10,922 genes whereas 28,762 genes were found to be expressed in the leaf transcriptome. Nine isolates were sufficient to sample 99% of the pangenome and its size did not increase with subsequent isolates, proving that it was closed (Table 1). Like *A. thaliana*, core genes were more expressed and more constrained than variable genes (Contreras-Moreira et al. 2017). Merging tissue-

specific transcriptomic and whole-genome sequencing data promises to become a powerful approach for future core- and pangenome determinations in metazoans and plants.

11 The Eukaryotic Pangenome

As François Jacob put it more than 40 years ago, gene evolution mainly deals with tinkering, molecular tinkering (Jacob 1977). Young eukaryotes (angiosperms, mammals) reshuffled gene exons and protein domains that already existed in old eukaryotes (fungi, excavata, monocellular animals, and algae), more than one billion years ago. There were very few real inventions after the first eukaryotes, some of them aforementioned here. An *Alu* element or a piece of a virus genome may be captured to make a new protein domain, transposons moved around, sometimes taking along a piece of DNA that would eventually become an exonic sequence, accumulation of mutations in a duplicated gene copy could ultimately create a new function by sub- or neofunctionalization. The redundant nature of eukaryotic genomes, particularly young ones, is only apparent. Eukaryotic core genes are hidden behind legions of transposons, successive whole-genome duplications and interspecific hybridizations, but one may ask how many genes are part of the eukaryotic core-genome. When trying to define it, exons or protein domains, rather than genes, should probably be considered as relevant genetic units, to circumvent issues due to molecular tinkering. Further definition of an eukaryotic pangenome will prove to be a long and complex task, but the accumulation of high-quality genome sequences and the exponential increase of computing power, might prove it to be a reachable goal in the forthcoming years.

In 2016, a German team tried to reconstitute the prokaryotic core-genome, using sequences from 1847 eubacteria and 134 archaebacteria species, covering 6.1 million protein-coding genes belonging to 286,000 families. They identified 355 proteins common to all species, that may be considered as the prokaryotic core-genome (Weiss et al. 2016, 2018). But one may ask whether this minimal set of core genes is sufficient to support life. In an attempt to create a hypothetical minimal genome, the J. Craig Venter institute applied synthetic genomics approaches to *Mycoplasma mycoides*. Using a combination of existing deletion data and literature mining, eight independent segments covering altogether the whole *M. mycoides* genome were synthesized. Each of these eight segments was individually reintroduced into bacteria, but only one of them produced a viable genome. Using high-throughput transposon mutagenesis, the team subsequently identified a set of 229 genes that would cause different levels of growth impairment. The eight DNA segments were rebuilt including these genes. Although each of the individual segment was able to produce a viable genome, addition of the eight segments in the same bacteria was lethal. Once the team eventually solved this synthetic lethality issue and succeeded in synthesizing a fully functional minimal genome, they discovered that the biological function of 146 genes (out of 473 encoded) could not be assigned. These genes

of unknown function were all needed to sustain *M. mycoides* life (Hutchison et al. 2016). This interesting work supports the conclusion that designing a minimal genome based on a core set of genes common to several isolates or to several species might not be sufficient to support life. Therefore, defining pangenome contents might prove essential to rewrite the genomes of more complex organisms, like eukaryotes.

As one last word, it must be noted that core-and pangenomes described here took only into consideration protein-coding genes. It is noteworthy that eukaryotes contain many more genes encoding various RNA species: tRNA, rRNA, snoRNA, scRNA, microRNA, and siRNA. Building the whole repertoire of such genes will be challenging but essential to define, at last, a complete eukaryotic pangenome.

Acknowledgments I wish to thank Héloïse Muller for useful suggestions and the Centre National de la Recherche Scientifique (CNRS) as well as the Institut Pasteur for their continuous support.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195

Anderson E (1949) Introgressive hybridization. Wiley, New York

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N et al (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature 444:171–178

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE (2004) Analysis of segmental duplications and genome assembly in the mouse. Genome Res 14:789–801

Bennett RJ, Johnson AD (2003) Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. EMBO J 22:2505–2515

Bodey GP, Mardani M, Hanna HA, Boktour M, Abbas J, Girgawy E, Hachem RY, Kontoyiannis DP, Raad I (2002) The epidemiology of *Candida glabrata* and *Candida albicans* fungemia in immunocompromised patients with cancer. Am J Med 112:380–385

Brosius J (1991) Retroposons – seeds of evolution. Science 251:753–753

Brugger J, Feulner G, Petri S (2017) Baby, it's cold outside: climate model simulations of the effects of the asteroid impact at the end of the Cretaceous. Geophys Res Lett 44(1):419–427. https://doi.org/10.1002/2016GL072241

Carradec Q, Pelletier E, Silva CD, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K et al (2018) A global ocean atlas of eukaryotic genes. Nat Commun 9:373

Carreté L, Ksiezopolska E, Pegueroles C, Gómez-Molero E, Saus E, Iraola-Guzmán S, Loska D, Bader O, Fairhead C, Gabaldón T (2018) Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. Curr Biol 28:15–27.e7

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S et al (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93

Contreras-Moreira B, Cantalapiedra CP, García-Pereira MJ, Gordon SP, Vogel JP, Igartua E, Casas AM, Vinuesa P (2017) Analysis of plant pan-genomes and transcriptomes with GET_HOMOLOGUES-EST, a clustering solution for sequences of the same species. Front Plant Sci 8:184

Cormack BP, Ghori N, Falkow S (1999) An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. Science 285:578–582

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Bescot NL, Probert I et al (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3:1700–1708

Deng C, Cheng C-HC, Ye H, He X, Chen L (2010) Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. PNAS 107:21593–21598

Domergue R, Castaño I, Peñas ADL, Zupancic M, Lockatell V, Hebel JR, Johnson D, Cormack BP (2005) Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. Science 308:866–870

Downie JA, Stewart JW, Brockman N, Schweingruber AM, Sherman F (1977) Structural gene for yeast iso-2-cytochrome c. J Mol Biol 113:369–384

Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218

Dujon B (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet 22:375–387

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E et al (2004) Genome evolution in yeasts. Nature 430:35–44

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. PNAS 99:16144–16149

Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res 22:908–924

Dunning Hotopp JC, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S et al (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–1756

Dupressoir A, Vernochet C, Bawa O, Harper F, Pierron G, Opolon P, Heidmann T (2009) Syncytin-A knockout mice demonstrate the critical role in placentation of a fusogenic, endogenous retrovirus-derived, envelope gene. PNAS 106:12127–12132

Emanuel BS, Shaikh TH (2001) Segmental duplications: an "expanding" role in genomic instability and disease. Nat Rev Genet 2:791–800

Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. Nature 440:623–630

Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. Science 334:1091–1097

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363–367

Fabre E, Muller H, Therizols P, Lafontaine I, Dujon B, Fairhead C (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. Mol Biol Evol 22:856–873

Fédry J, Liu Y, Péhau-Arnaudet G, Pei J, Li W, Tortorici MA, Traincard F, Meola A, Bricogne G, Grishin NV et al (2017) The ancient gamete fusogen HAP2 is a eukaryotic class II fusion protein. Cell 168:904–915.e10

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates, Sunderland, MA

Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, Hejnol A, Henrissat B, Koszul R, Aury J-M et al (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500:453–457

Fouts DE, Brinkac L, Beck E, Inman J, Sutton G (2012) PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res 40:e172–e172

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, Kuo A, Paredez A, Chapman J, Pham J et al (2010) The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell 140:631–642

Gabaldón T, Fairhead C (2019) Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. Curr Genet 65(1):93–98

Gabaldon T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R et al (2013) Comparative genomics of emerging pathogens in the *Candida glabrata* clade. BMC Genomics 14:623

Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100:659–674

Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive horizontal gene transfer in bdelloid rotifers. Science 320:1210–1213

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al (1996) Life with 6000 genes. Science 274:546–567

Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D et al (1997) The yeast genome directory. Nature 387 (suppl):1–105

Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP et al (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun 7:13390

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y et al (2010) A draft sequence of the neandertal genome. Science 328:710–722

Guimarães LC, Florczak-Wyspianska J, de Jesus LB, Viana MVC, Silva A, Ramos RTJ, Soares S d C, Soares S d C (2015) Inside the pan-genome – methods and software overview. Curr Genomics 16:245–252

Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716

Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, Zeng Q, Zisson E, Wang JM, Greenberg JM et al (2015) Genetic and phenotypic intra-species variation in *Candida albicans*. Genome Res 25:413–425

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K et al (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, Liang Y, Yi X, He M, Somel M et al (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512:194–197

Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L et al (2016) Design and synthesis of a minimal bacterial genome. Science 351:aad6253–aad6253

i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104:595–600

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, Harris D, Aslett M, Barrell JF, Butler G, Citiulo F et al (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. Genome Res 19:2231–2244

Jacob F (1977) Evolution and tinkering. Science 196:1161–1166

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A et al (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946–957

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Jiang Z, Tang H, Ventura M, Cardone MF, Marques-Bonet T, She X, Pevzner PA, Eichler EE (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat Genet 39:1361–1368

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT et al (2004) The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci USA 101:7329–7334

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428:617–624

Koszul R, Caburet S, Dujon B, Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. EMBO J 23:234–243

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J et al (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63

Li Y, Zhou G, Ma J, Jiang W, Jin L, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L et al (2014) *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32:1045–1052

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJT, van Oudenaarden A, Barton DBH, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ (2009) Population genomics of domestic and wild yeasts. Nature 458 (7236):337–341

López-García P, Moreira D (2006) Selective forces for the origin of the eukaryotic nucleus. BioEssays 28:525–533

Makałowski W, Mitchell GA, Labuda D (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet 10:188–193

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC et al (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527–1541

Mereschowsky K (1999) On the nature and origin of chromatophores in the plant kingdom. Eur J Phycol 34:287–295

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? PLoS Biol 9:e1001127

Morales L, Dujon B (2012) Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. Microbiol Mol Biol Rev 76:721–739

Moran NA, Jarvik T (2010) Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science 328:624–627

Muller H, Thierry A, Coppée J-Y, Gouyette C, Hennequin C, Sismeiro O, Talla E, Dujon B, Fairhead C (2009) Genomic polymorphism in the population of *Candida glabrata*: gene copy-number variation and chromosomal translocations. Fungal Genet Biol 46(3):264–267. https://doi.org/10.1016/j.fgb.2008.11.006

Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S et al (2015) Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun 6:8018

Nguyen H-V, Legras J-L, Neuvéglise C, Gaillardin C (2011) Deciphering the hybridisation history leading to the lager lineage based on the mosaic genomes of *Saccharomyces bayanus* strains NBRC1948 and CBS380T. PLoS One 6:e25821

Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I (2014) The genome portal of the department of energy joint genome institute: 2014 updates. Nucleic Acids Res 42:D26–D31

Normile D, 2017, and Am, 8:00 (2017) Plant scientists plan massive effort to sequence 10,000 genomes

Nurminsky DI, Nurminskaya MV, Aguiar DD, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396:572–575

Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, Town CD, Nixon J, Krishnakumar V, Bidwell SL et al (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. Genome Biol 15:R77

Payen C, Koszul R, Dujon B, Fischer G (2008) Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. PLoS Genet 5: e1000175

Pennisi E (2016) Shaking up the tree of life. Science 354:817–821

Pennisi E (2017) Biologists propose to sequence the DNA of all life on Earth. Science Magazine, Feb 27, 2017

Peter J, Chiara MD, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A et al (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 556:339–344

Pfaller MA, Diekema DJ (2004) Twelve years of fluconazole in clinical practice: global trends in species distribution and fluconazole susceptibility of bloodstream isolates. Clin Microbiol Infect 10:11–23

Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, Zaina G, Bastien C, Cattonaro F, Marroni F et al (2016) Characterization of the poplar pan-genome by genome-wide identification of structural variation. Mol Biol Evol 33:2706–2719

Polakova S, Blume C, Zarate JA, Mentel M, Jorck-Ramberg D, Stenderup J, Piskur J (2009) Formation of new chromosomes as a virulence mechanism in yeast *Candida glabrata*. Proc Natl Acad Sci USA 106:2688–2693

Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C et al (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, Mayer C, Miller J, Monier A, Salamov A et al (2013) Pan genome of the phytoplankton *Emiliania* underpins its global distribution. Nature 499:209–213

Renne PR, Sprain CJ, Richards MA, Self S, Vanderkluysen L, Pande K (2015) State shift in Deccan volcanism at the Cretaceous-Paleogene boundary, possibly induced by impact. Science 350:76–78

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI et al (2011) Comparative functional genomics of the fission yeasts. Science 332:930–936

Richard G-F, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727

Rolland T, Neuvéglise C, Sacerdot C, Dujon B (2009) Insertion of horizontally transferred genes within conserved syntenic regions of yeast genomes. PLoS One 4:e6515

Rolland T, Dujon B, Richard GF (2010) Dynamic evolution of megasatellites in yeasts. Nucleic Acids Res 38:4731–4739

Sagan L (1967) On the origin of mitosing cells. J Theor Biol 14:225–IN6

Schacherer J, Tourrette Y, Souciet J-L, Potier S, de Montigny J (2004) Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae*. Genome Res 14:1291–1297

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Seki R, Li C, Fang Q, Hayashi S, Egawa S, Hu J, Xu L, Pan H, Kondo M, Sato T et al (2017) Functional roles of Aves class-specific *cis*-regulatory elements on macroevolution of bird-specific features. Nat Commun 8:14229

Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S et al (2018) The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561:113–116

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U et al (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature 466:720–726

Tekaia F, Dujon B, Richard G-F (2013) Detection and characterization of megasatellites in orthologous and nonorthologous genes of 21 fungal genomes. Eukaryot Cell 12:794–803

Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, Xie C et al (2016) Deep sequencing of 10,000 human genomes. PNAS 113:11901–11906

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". PNAS 102:13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68–74

The 1001 Genomes Consortium (2016) 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491

Thierry A, Bouchier C, Dujon B, Richard G-F (2008) Megasatellites: a peculiar class of giant minisatellites in genes involved in cell adhesion and pathogenicity in *Candida glabrata*. Nuclic Acids Res 36:5970–5982

Thierry A, Dujon B, Richard G-F (2009) Megasatellites: a new class of large tandem repeats discovered in the pathogenic yeast *Candida glabrata*. Cell Mol Life Sci 67:671–676

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) The sequence of the human genome. Science 291:1304–1351

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in Arabidopsis. Science 290:2114–2117

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60

Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, Martin WF (2016) The physiology and habitat of the last universal common ancestor. Nat Microbiol 1:16116

Weiss MC, Preiner M, Xavier JC, Zimorski V, Martin WF (2018) The last universal common ancestor between ancient Earth chemistry and the onset of genetics. PLoS Genet 14:e1007518

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708–713

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV et al (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Science 324:268–272

Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16:187
Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU et al (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353–358
Zhang G (2015) Genomics: bird sequencing project takes off. Nature 522:34
Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278–284