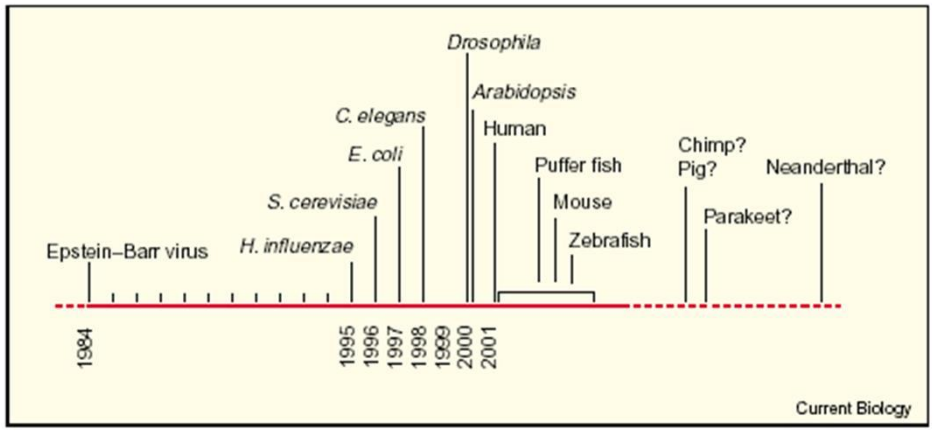
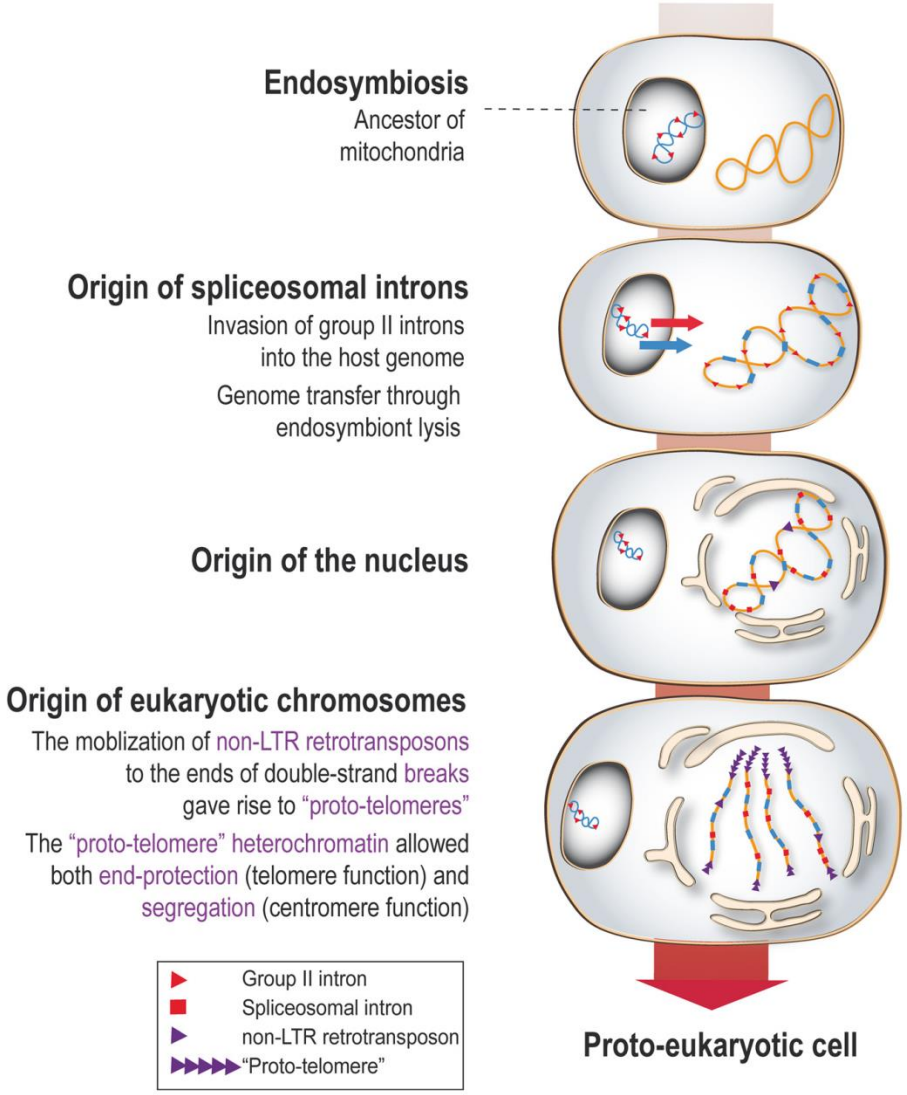
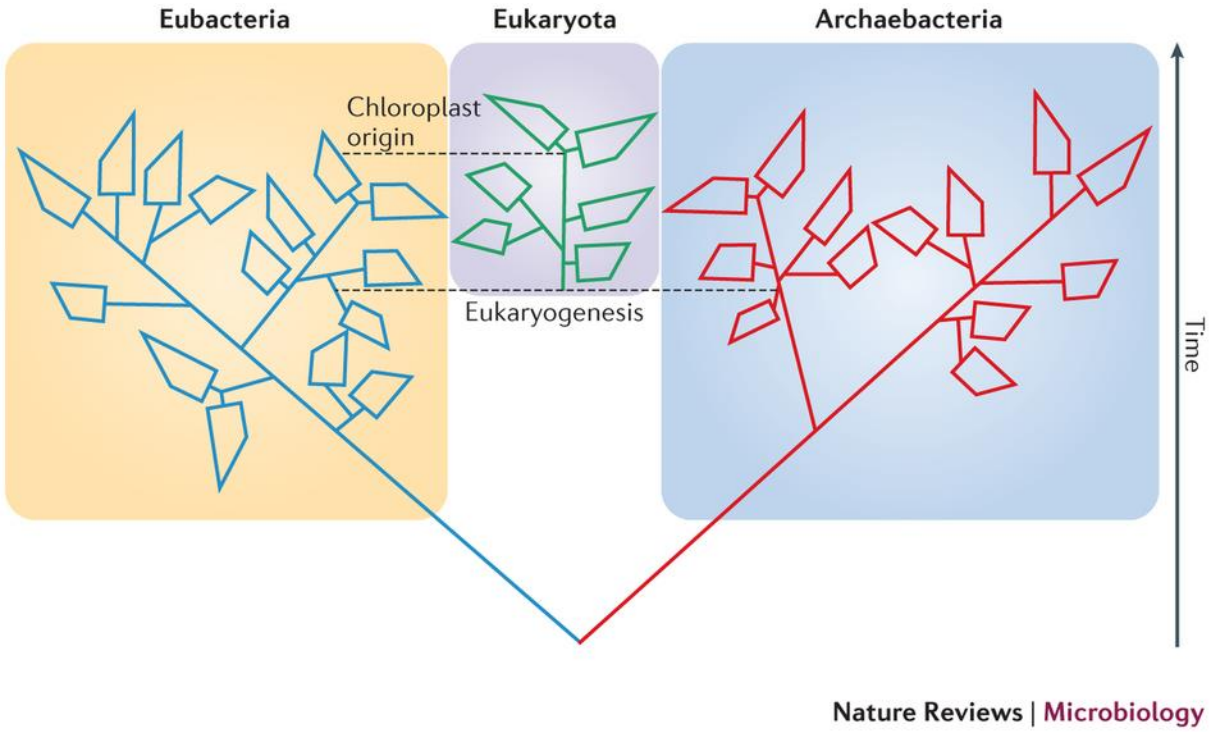


Past, present, future: a timeline marking a small sampling of completed, ongoing, and proposed genome initiatives. From diverse mitochondria to the nuclear genomes of many model organsms, sequence information from over 800 genomes is available on the internet.

PR13_Eukariontska genomika

Kompleksnost eukariontskih genomov





The ring of life hypothesis

Schematic representation of the flow of genetic material from the two major prokaryotic groups into the base of the eukaryotes and the separate flow of genetic material from cyanobacteria into plastid-containing eukaryotes.

Elements of eukaryotic genomes (nuclear)

Chromosomes: linear, centromeres, telomeres, origins of replication, replicons

Protein-coding genes and **spliceosomal introns**

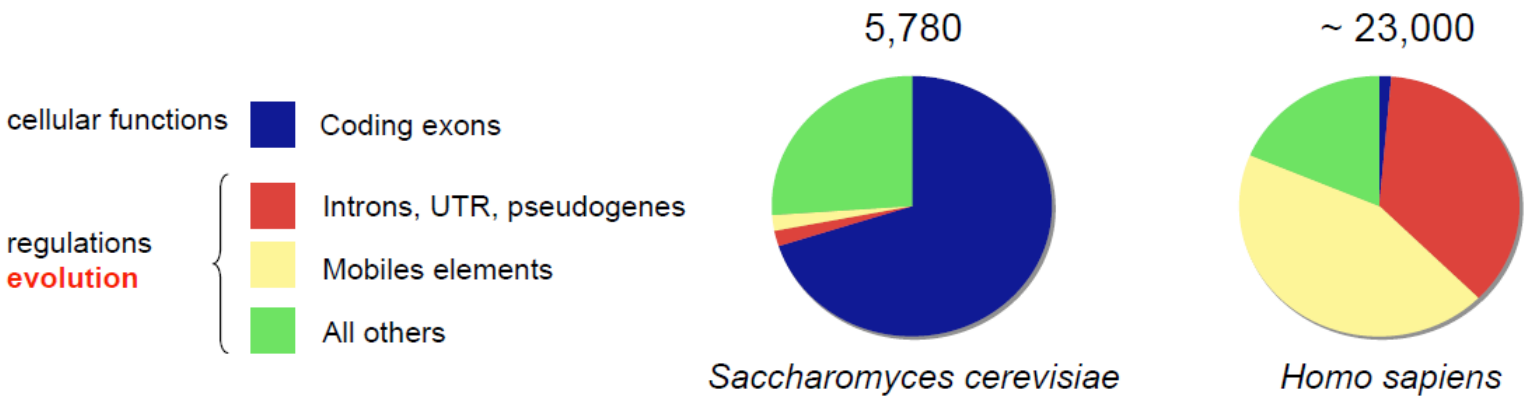
Genes for non coding RNAs: rRNAs, tRNAs, snoRNAs, snRNAs, microRNAs .....

Mobile genetic elements: and their remnants

Pseudogenes: and processed pseudogenes

Satellite DNAs: micro-, minisatellites, repeated sequences

Fragments of organellar DNAs: NUMTs and NUPTs



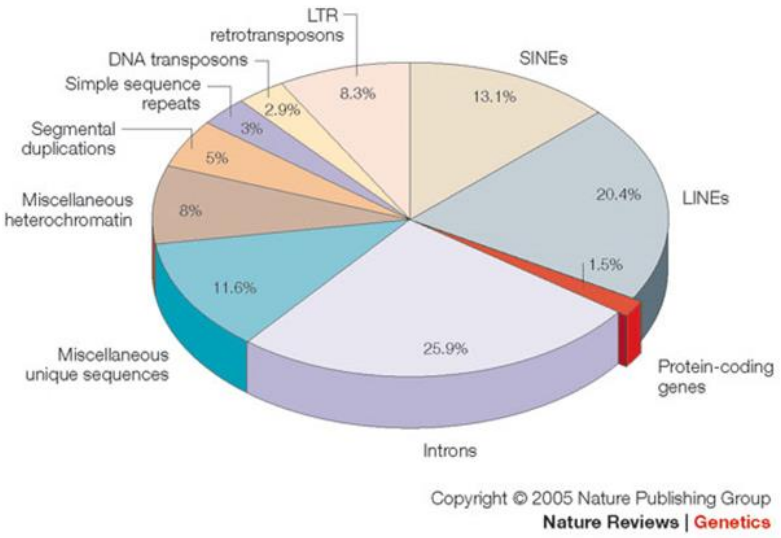
Duplicated genes and regions, acquired DNA sequences, newly-created genes :

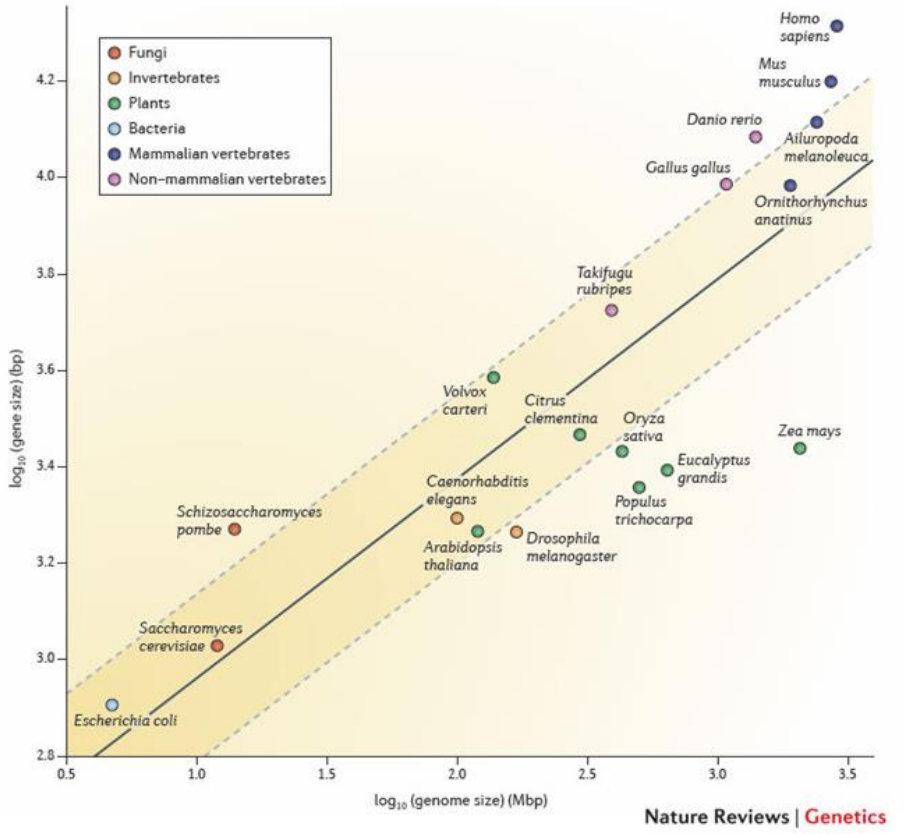
> intense dynamics of genome modification and evolution

> genomes contain dispersed repeated sequences

pb of sequence assembly and interpretation

Major component of eukaryotic genomes are transposable elements

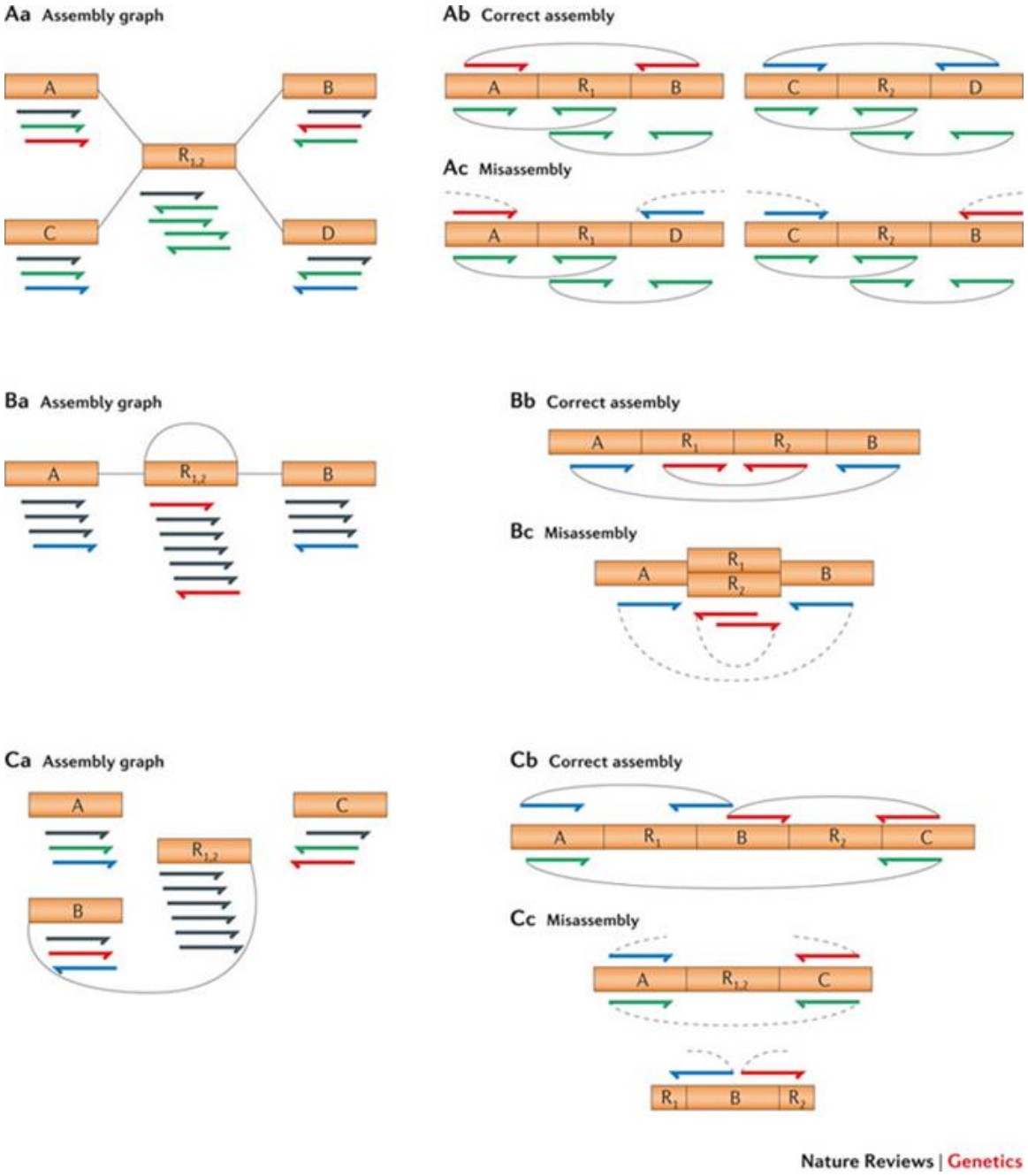




Genome and gene sizes for a representative set of genomes

Gene size is plotted as a function of genome size for some representative bacteria, fungi, plants and animals. This figure illustrates a simple rule of thumb: in general, bigger genomes have bigger genes. Thus, accurate annotation of a larger genome requires a more contiguous genome assembly in order to avoid splitting genes across scaffolds. Note too that although the human and mouse genomes deviate from the simple linear model shown here, the trend still holds. Their unusually large genes are likely to be a consequence of the mature status of their annotations, which are much more complete as regards annotation of alternatively spliced transcripts and untranslated regions than those of most other genomes.

Težave pri sekveniranju in assemblyu eukariontskih genomov



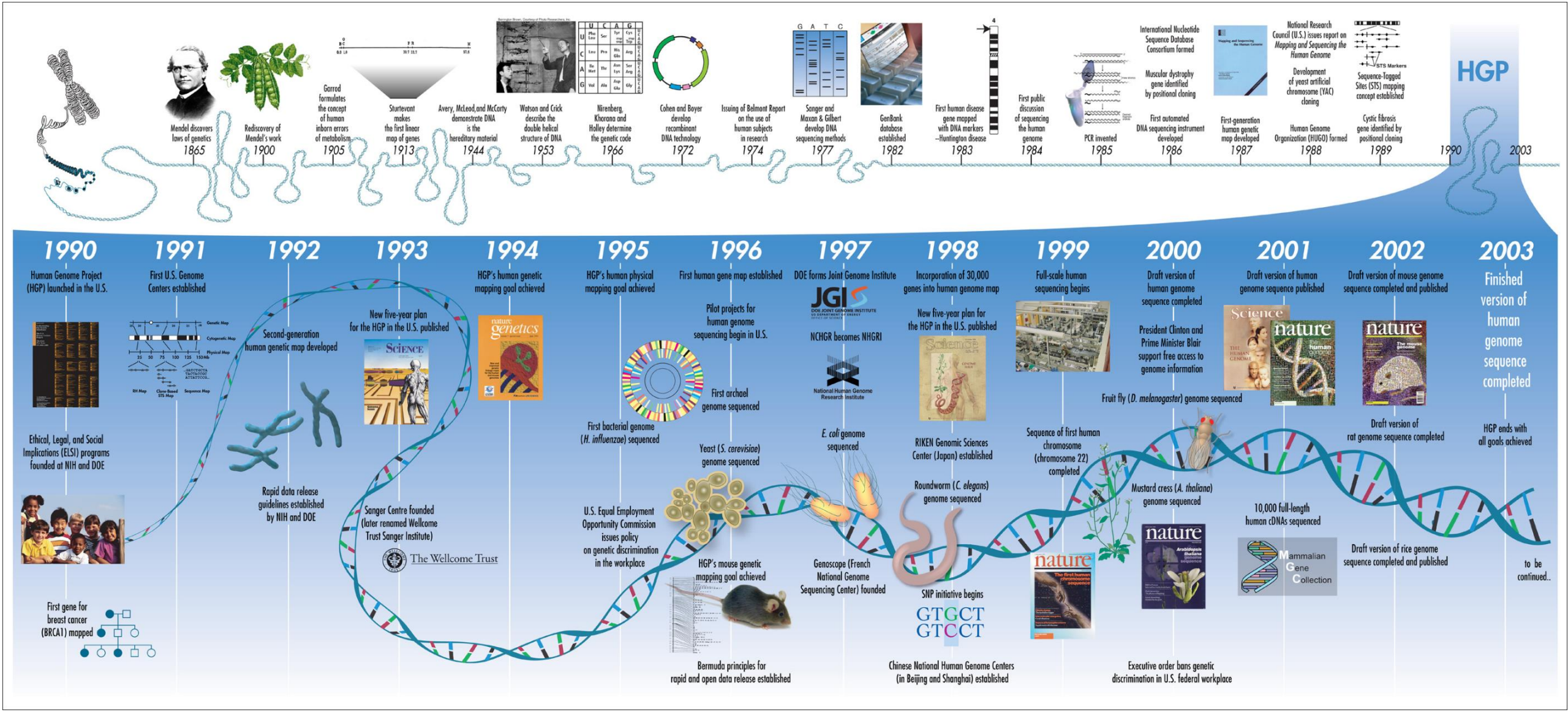
Assembly errors caused by repeats

A | Rearrangement assembly error caused by repeats. Aa | An example assembly graph involving six contigs, two of which are identical (R1 and R2). The arrows shown below each contig represent the reads that are aligned to it. Ab | The true assembly of two contigs, showing mate-pair constraints for the red, blue and green paired reads. Ac | Two incorrectly assembled chimeric contigs caused by the repetitive regions R1 and R2. Note that all reads align perfectly to the misassembled contigs, but the mate-pair constraints are violated.

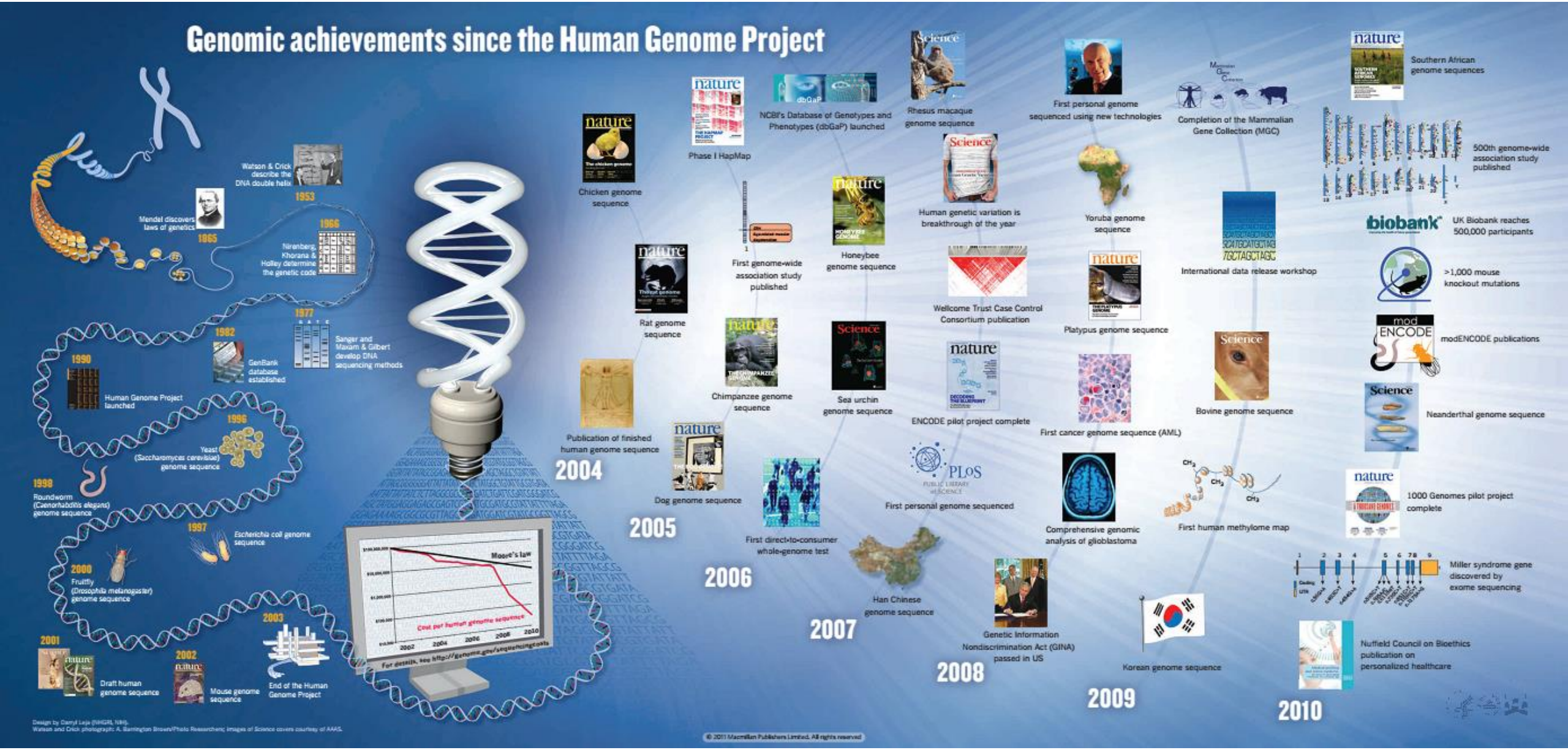
B | A collapsed tandem repeat. Ba | The assembly graph contains four contigs, where R1 and R2 are identical repeats. Bb | The true assembly, showing mate-pair constraints for the red and blue paired reads, which are oriented correctly and spaced the correct distance apart. Bc | A misassembly that is caused by collapsing repeats R1 and R2 on top of each other. Read alignments remain consistent, but mate-pair distances are compressed. A different misassembly of this region might reverse the order of R1 and R2.

C | A collapsed interspersed repeat. Ca | The assembly graph contains five contigs, where R1 and R2 are identical repeats. Cb | In the correct assembly, R1 and R2 are separated by a unique sequence. Cc | The two copies of the repeat are collapsed onto one another. The unique sequence is then left out of the assembly and appears as an isolated contig with partial repeats on its flank.

Zgodovina sekveniranja eukariontskih genomov I



Zgodovina sekveniranja eukariontskih genomov II

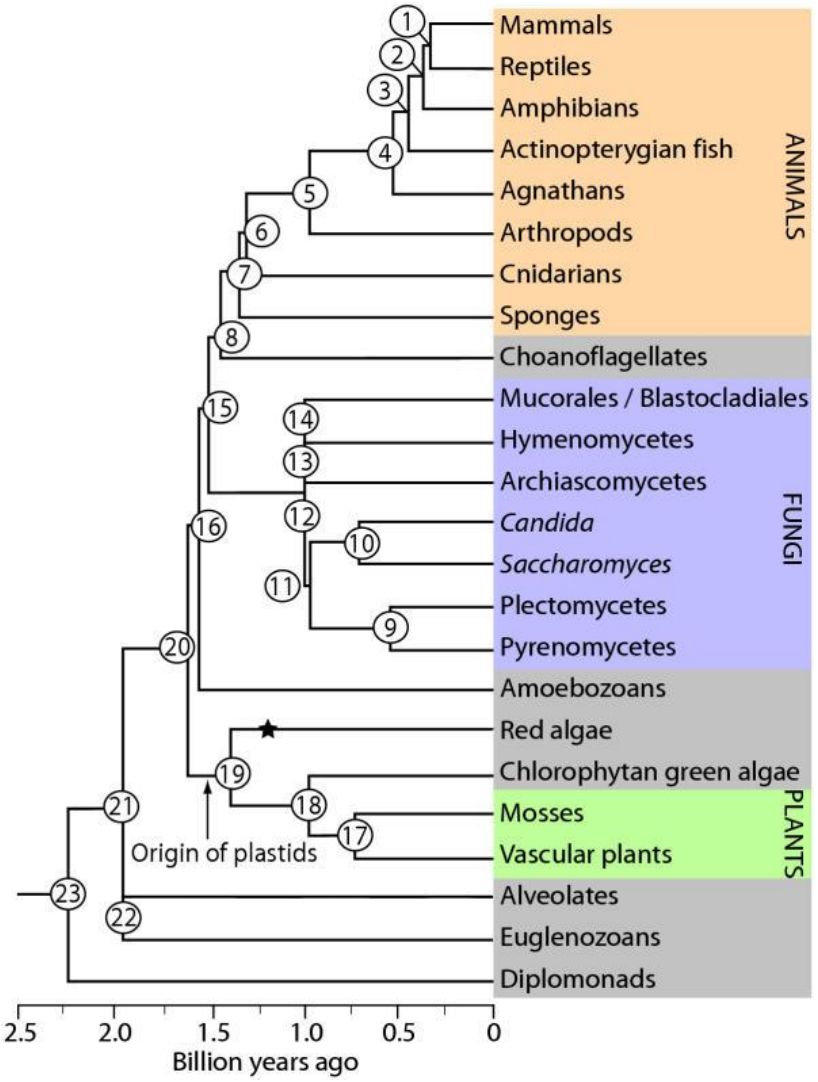


GOLD Eukaryotic genome projects (20.5. 2014):
total - 8143,
genomes - 5562,
transcriptomes – 1044,
resequencing – 1387,
uncultured – 26

Complete: 2584,
Permanent draft: 859,
Draft: 643,
In progress: 4563,
Targeted: 6.

Fungi: 3224 genomes
Metazoa: 1702
Plants: 1975
„Protists“: 570

A timescale of eukaryote evolution



The world of eukaryotes from genomics (P. Keeling, 2005)

Green algae

Plants, Ferns, Moss

Red algae

Apicomplexa (*Plasmodium*)

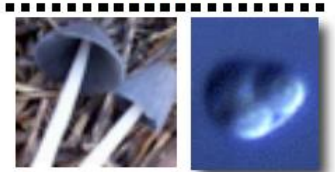
Ciliates (*Paramecium*)

Diatoms

Brown algae

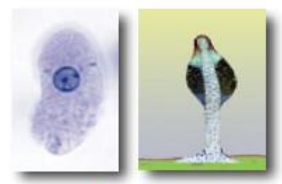
Trypanosoma

Foraminifera

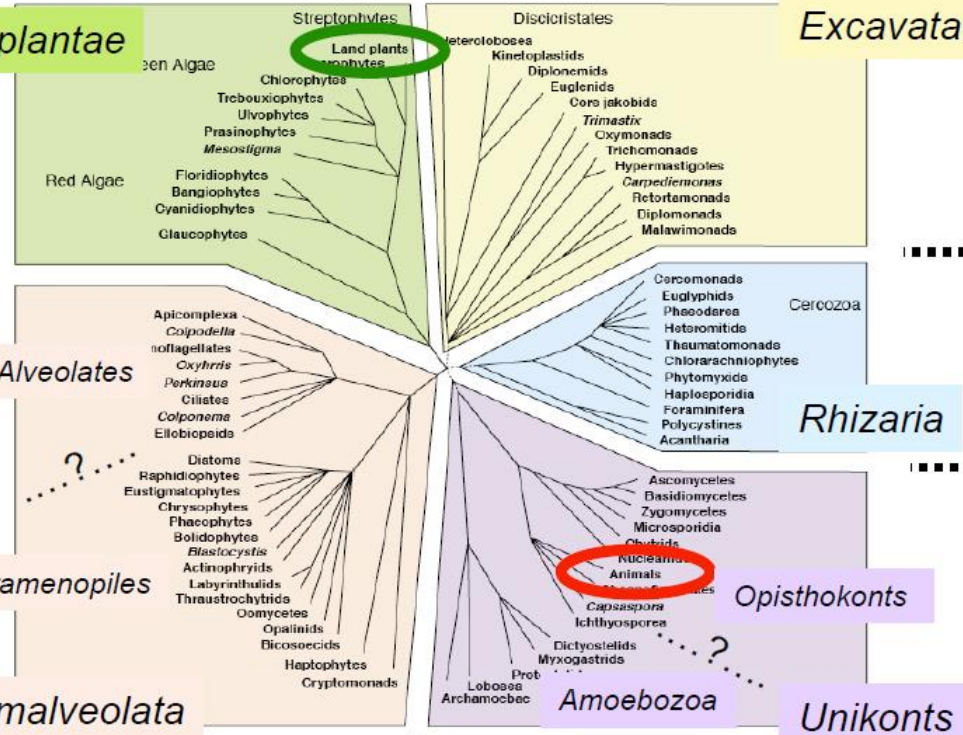


Fungi, yeasts

Animal world

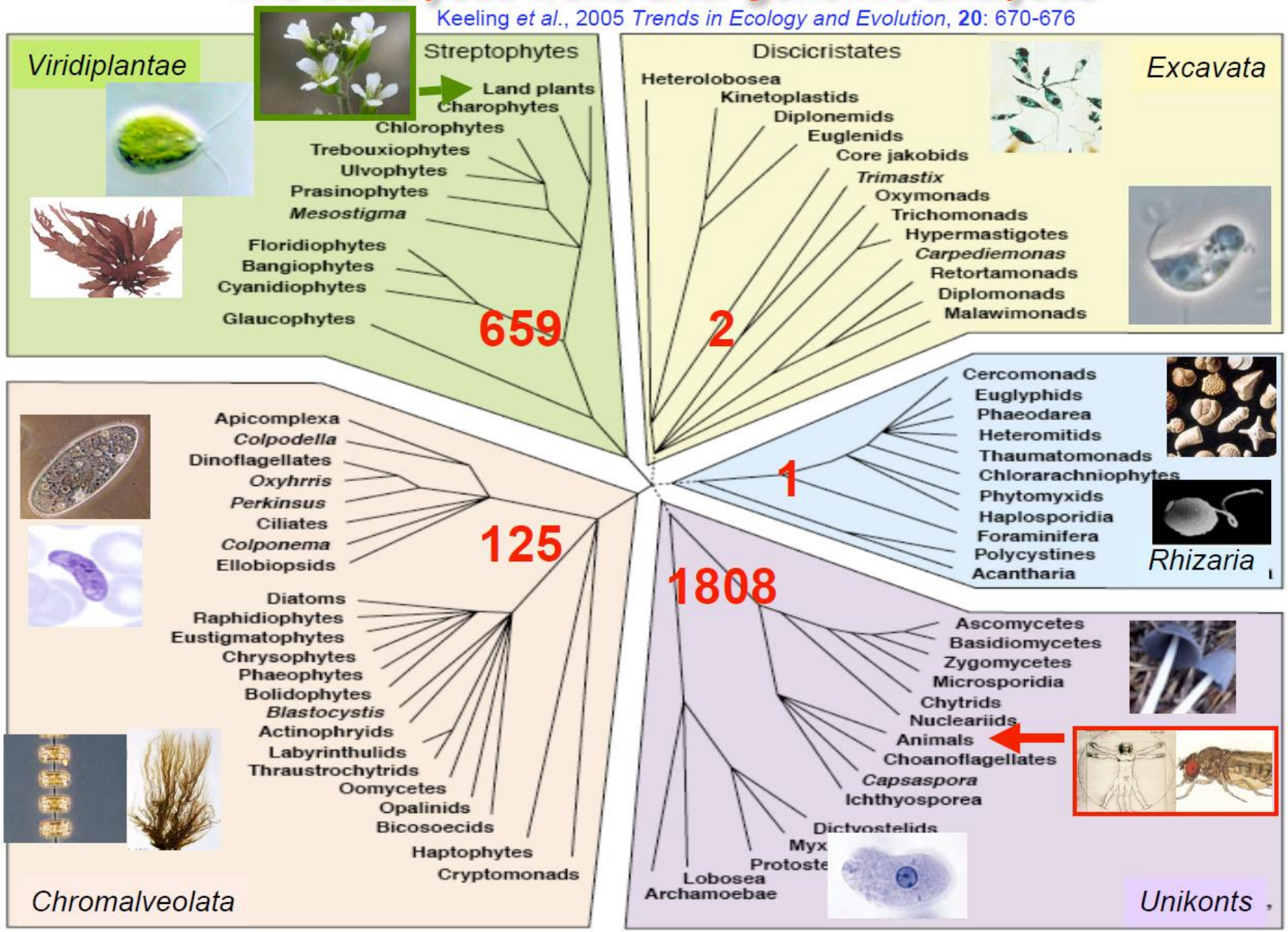


Amoebas

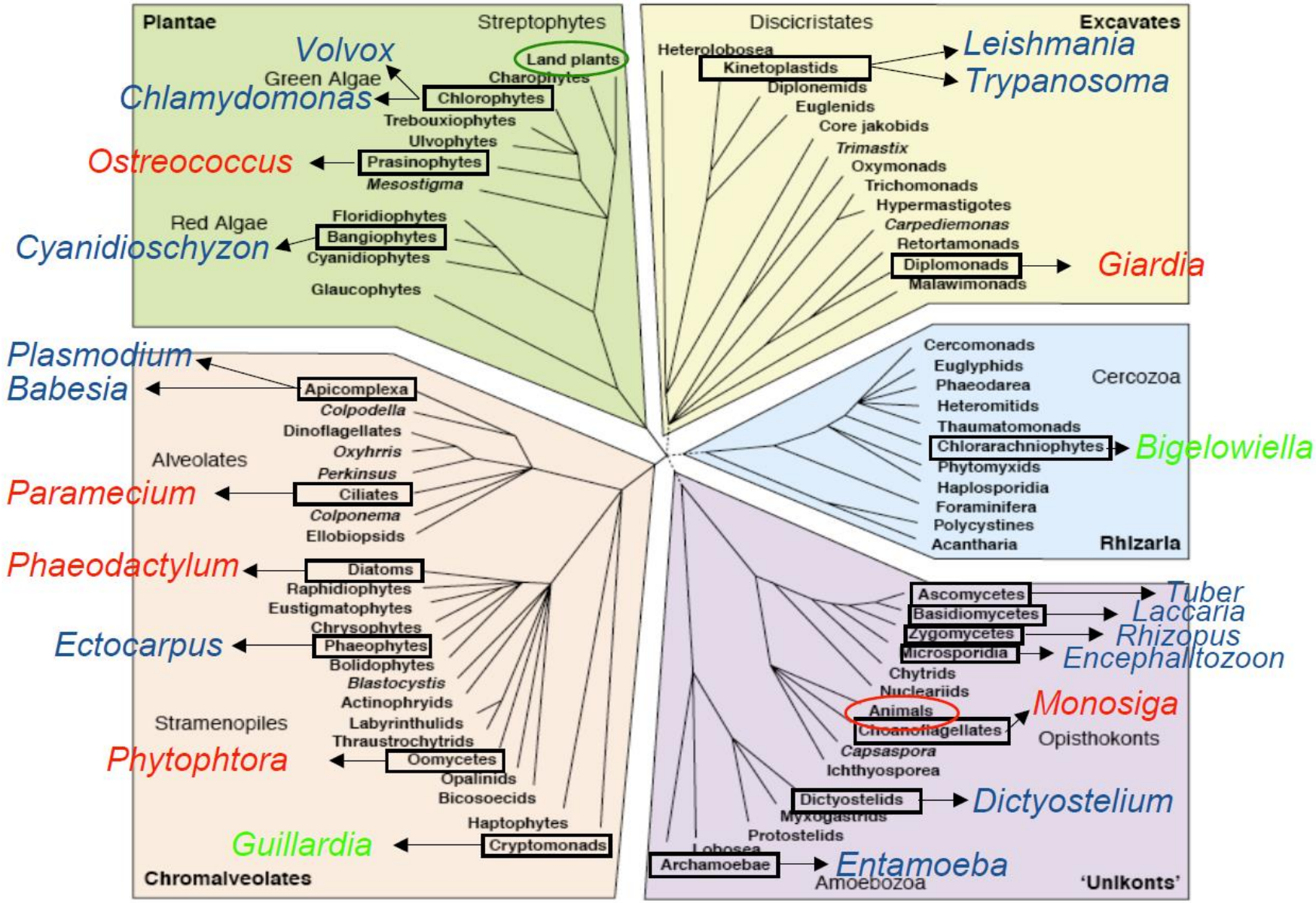


The eukaryotic world after genomic analyses

Keeling *et al.*, 2005 *Trends in Ecology and Evolution*, **20**: 670-676

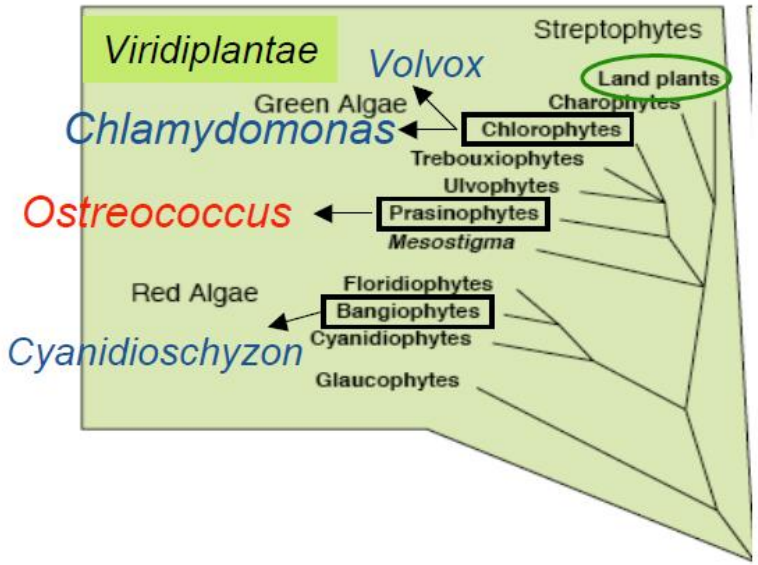


Genomes of unicellular eukaryotes



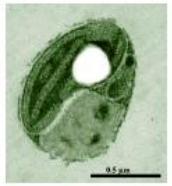
Keeling *et al.*, 2005 *Trends in Ecology and Evolution*, **20**: 670-676

Genomes of unicellular eukaryotes

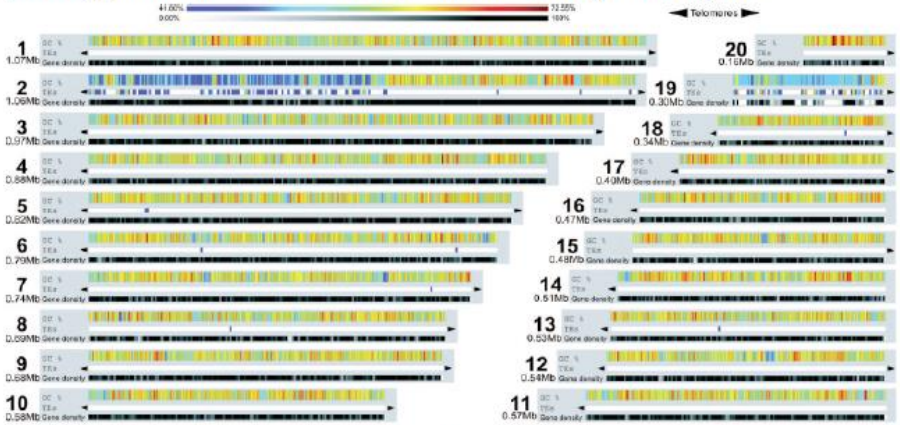


The genome of *Ostreococcus tauri*

Derelle *et al.*, 2006 *PNAS* **103**: 11647-11652



aim: genome of the smallest free living eukaryote



The genome of *Ostreococcus lucimarinus*

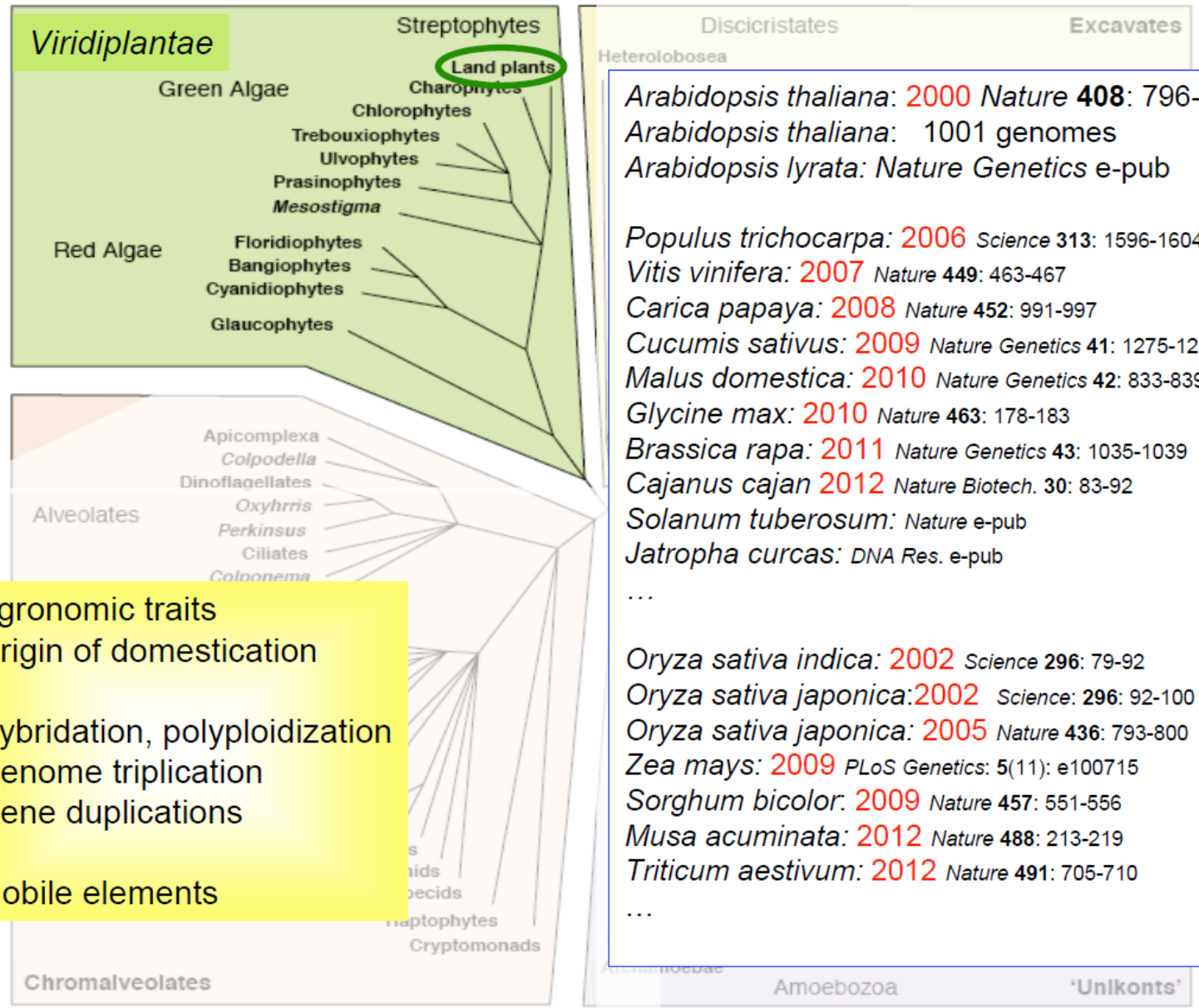
Palenik *et al.*, 2007 PNAS **104**: 7705-7710

- Genome size: 12.6 Mb, 20 chromosomes
- **Compositional heterogeneity** related to transposons (includes a 146 kb-long **segmental duplication**)

	O. lucimarinus	*O. tauri*
genome size (Mb)	13.2	12.6
chromosomes	21	20
protein-coding genes	7651	7892
split genes (%)	20	25

- Multiple mechanisms contribute to species divergence, act differently on different chromosomes.
- **Horizontal gene transfer** altering cell-surface characteristics.
- Numerous **gene fusions** 330 (*O.t.*), 348 (*O. l.*) of which 137 are common to both species.
- Numerous (20) genes for selenocysteine-containing proteins (TGA codons).

Genomes of *Streptophyta*



Arabidopsis thaliana: 2000 *Nature* **408**: 796-815
Arabidopsis thaliana: 1001 genomes
Arabidopsis lyrata: *Nature Genetics* e-pub

Populus trichocarpa: 2006 *Science* **313**: 1596-1604
Vitis vinifera: 2007 *Nature* **449**: 463-467
Carica papaya: 2008 *Nature* **452**: 991-997
Cucumis sativus: 2009 *Nature Genetics* **41**: 1275-1283
Malus domestica: 2010 *Nature Genetics* **42**: 833-839
Glycine max: 2010 *Nature* **463**: 178-183
Brassica rapa: 2011 *Nature Genetics* **43**: 1035-1039
Cajanus cajan 2012 *Nature Biotech.* **30**: 83-92
Solanum tuberosum: *Nature* e-pub
Jatropha curcas: *DNA Res.* e-pub
…

Oryza sativa indica: 2002 *Science* **296**: 79-92
Oryza sativa japonica:2002 *Science*: **296**: 92-100
Oryza sativa japonica: 2005 *Nature* **436**: 793-800
Zea mays: 2009 *PLoS Genetics*: **5**(11): e100715
Sorghum bicolor: 2009 *Nature* **457**: 551-556
Musa acuminata: 2012 *Nature* **488**: 213-219
Triticum aestivum: 2012 *Nature* **491**: 705-710
…

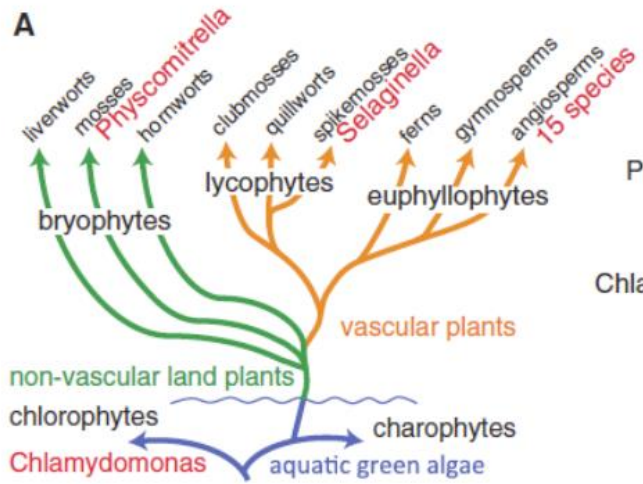
Agronomic traits
Origin of domestication

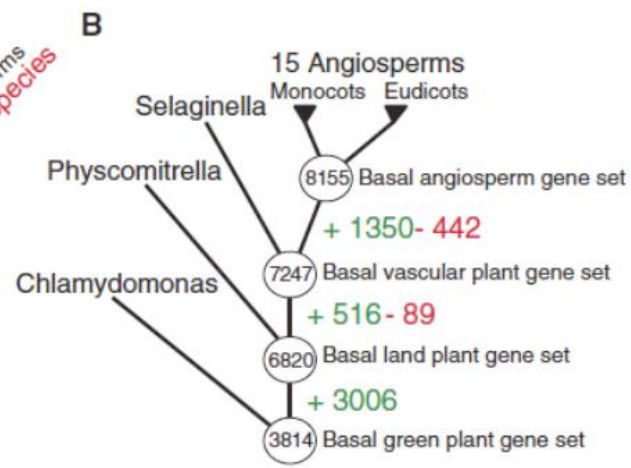
Hybridation, polyploidization
Genome triplication
Gene duplications

Mobile elements

The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants

Stefan A. Rensing,[1] Daniel Lang,[1] Andreas D. Zimmer,[1] Astrid Terry,[2] Asaf Salamov,[3] Harris Shapiro,[3] Tomoaki Nishiyama,[4] Pierre-Francois Perroud,[5] Erika A. Lindquist[3]



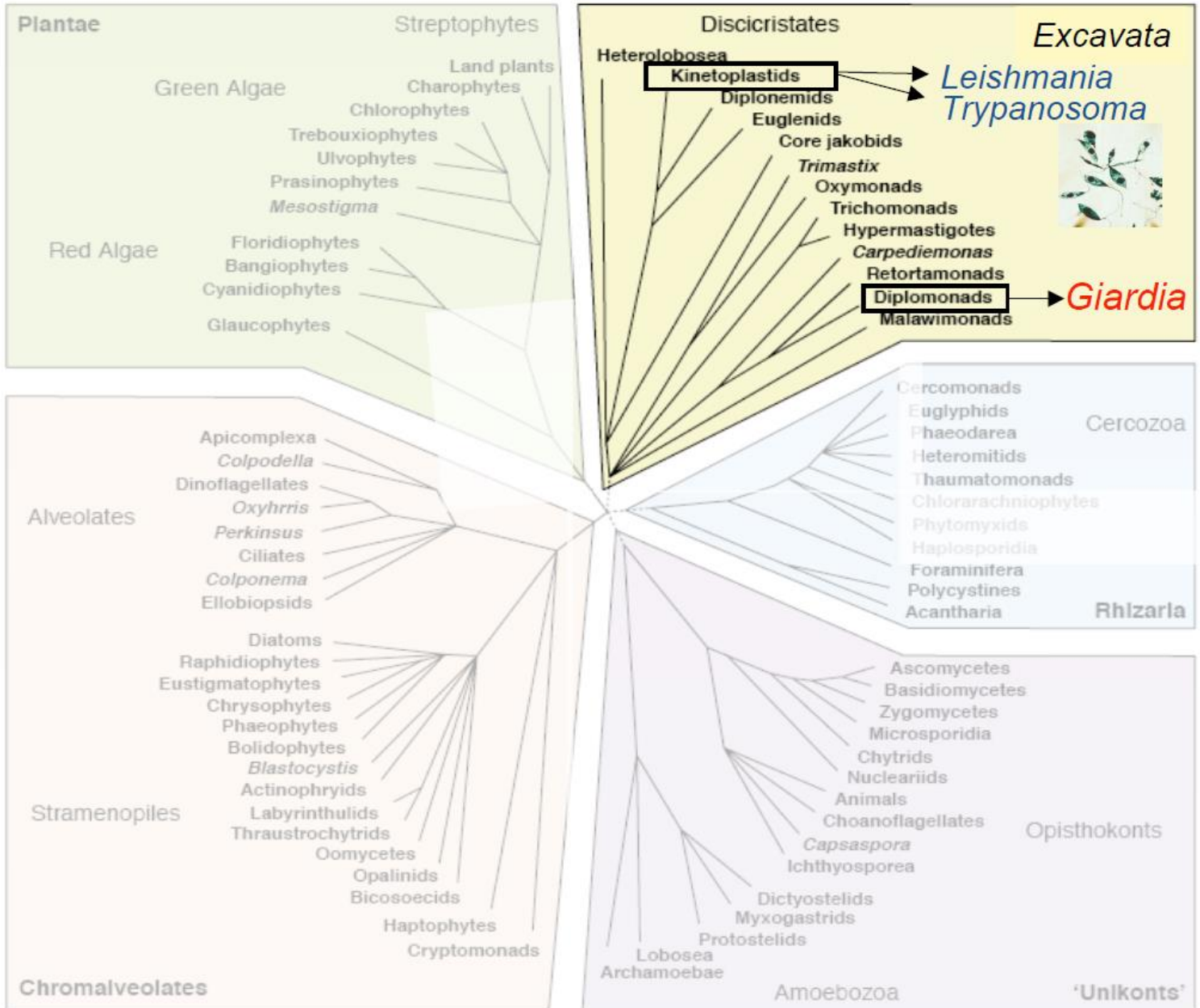


Minimal numbers of gene families in ancestors.
Gene family gains (+) and loss (-)

The Selaginella Genome Identifies Genetic Changes Associated with the Evolution of Vascular Plants

Jo Ann Banks,[1*] Tomoaki Nishiyama,[2,3] Mitsuyasu Hasebe,[3,4,5] John L. Bowman,[6,7] Michael Gribskov,[8] Claude dePamphilis,[9,10,11] Victor A. Albert,[12] Naoki Aono,[4] Tsuyoshi Aoyama,[4,5]

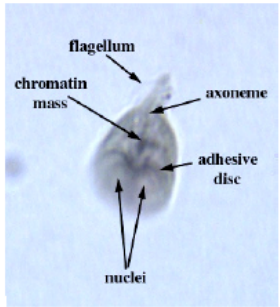
Genomes of unicellular eukaryotes: *Excavata*



Genomes of unicellular eukaryotes: *Excavata*

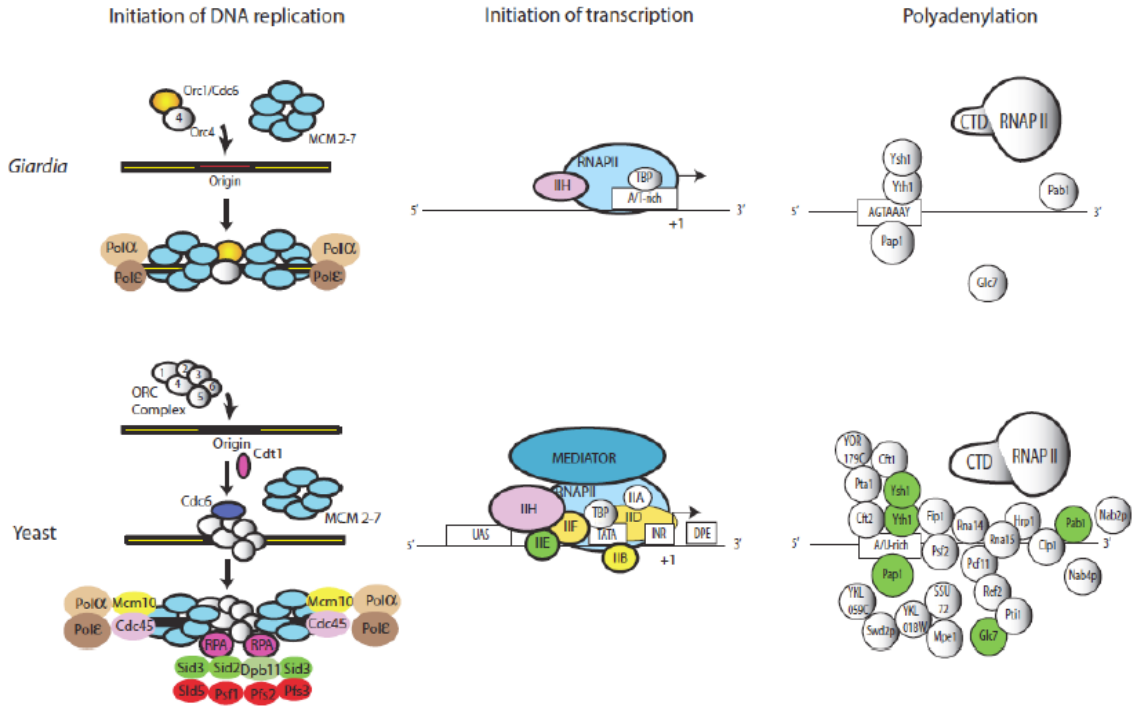
The genome of *Giardia lamblia*

Morrison *et al.*, 2007 *Science* **317**:1921-1926 (draft sequence 92 scaffolds)



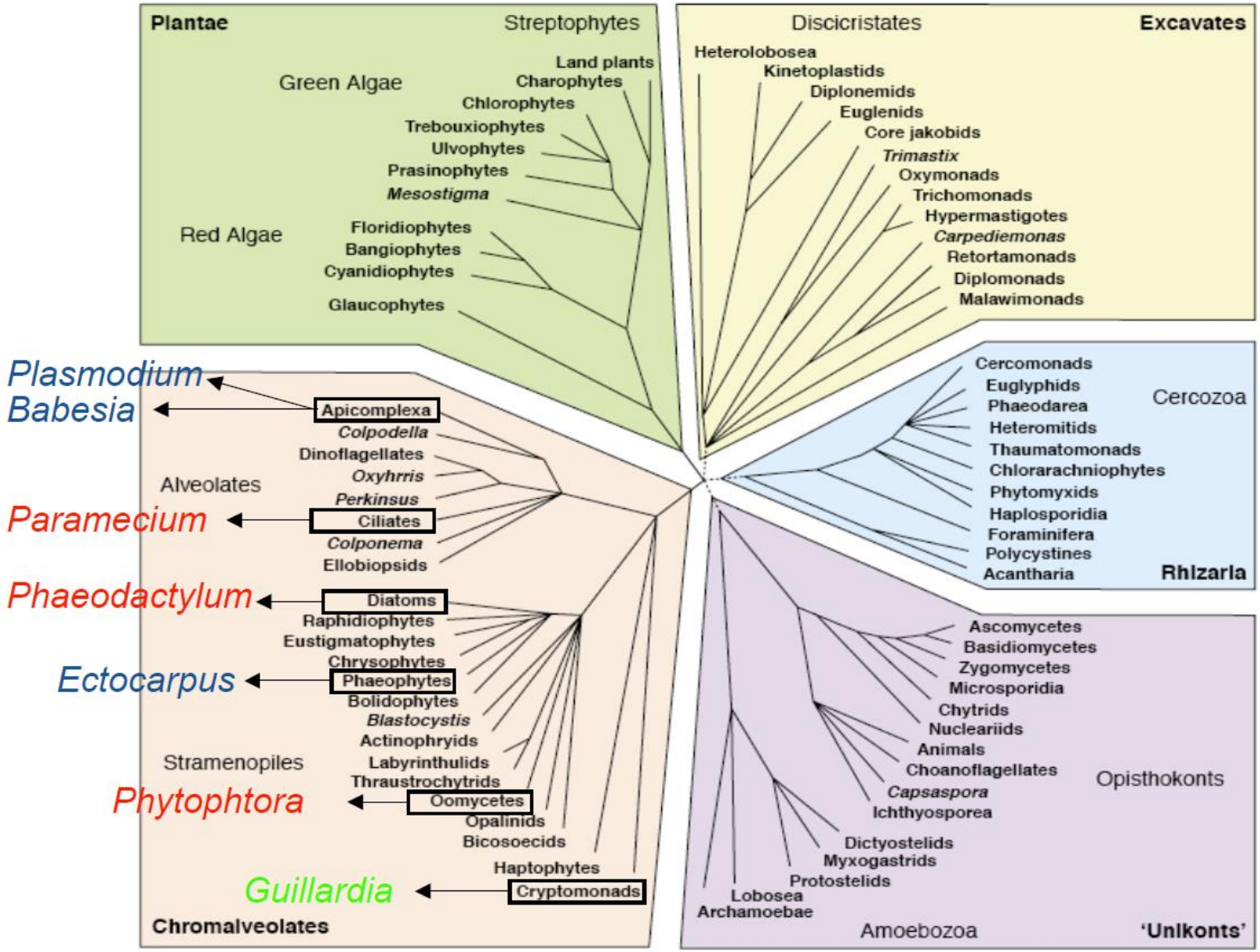
human intestinal parasite, flagellated trophozoites attach to epithelial cells
two diploid nuclei, no mitochondria, no peroxisomes

- Genome ~11.7 Mb, 5 chromosomes, 6470 annotated CDS, very few introns (4)
- Low degree of heterozygosity (0.01% between the 4 genomes)



- Simplified molecular machinery, cytoskeletal structure and metabolic pathways.
- Either early divergence or **regressive evolution.**
- Frequent **insertion of motifs** (up to 101 amino-acids) in conserved proteins.
- Numerous traces of horizontal gene acquisitions

Genomes of unicellular eukaryotes



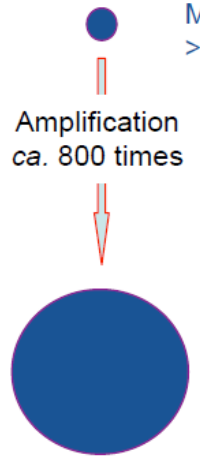
Keeling *et al.*, 2005 *Trends in Ecology and Evolution*, **20**: 670-676

Genomes of unicellular eukaryotes

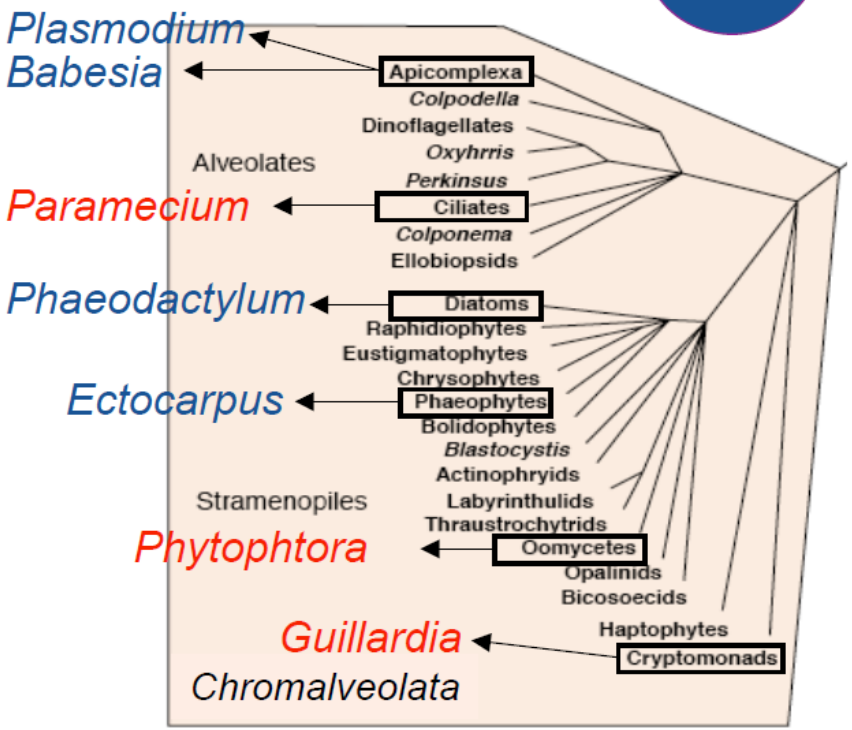
The macronuclear genome of *Paramecium tetraurelia*

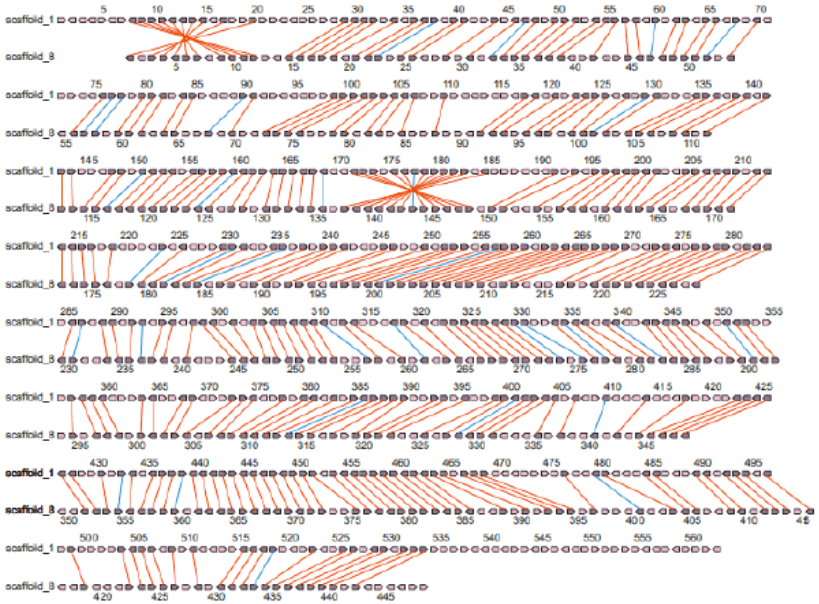
Aury *et al.*, (2006) *Nature* **444**: 171-178

697 scaffolds, totalling 72 Mb



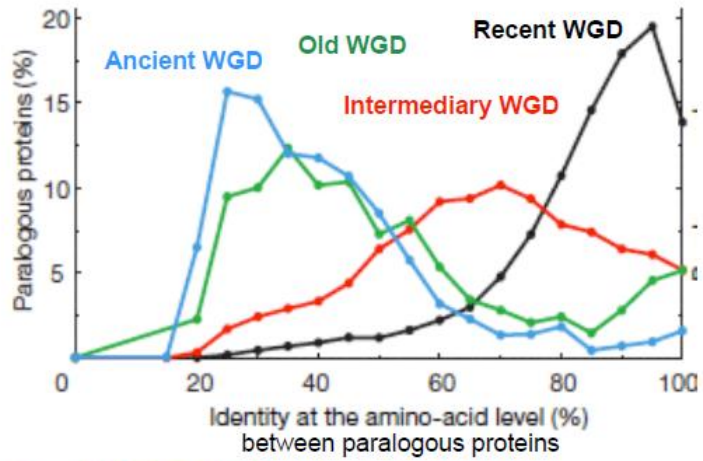
- Precise elimination of > 10000 short, unique copy elements ---> **Reconstruction of functional genes**
- Imprecise elimination of transposable elements and other repeated sequences ---> **Chromosome fragmentation, *de novo* telomere addition, internal deletions**

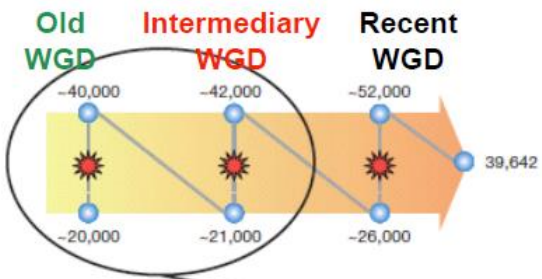


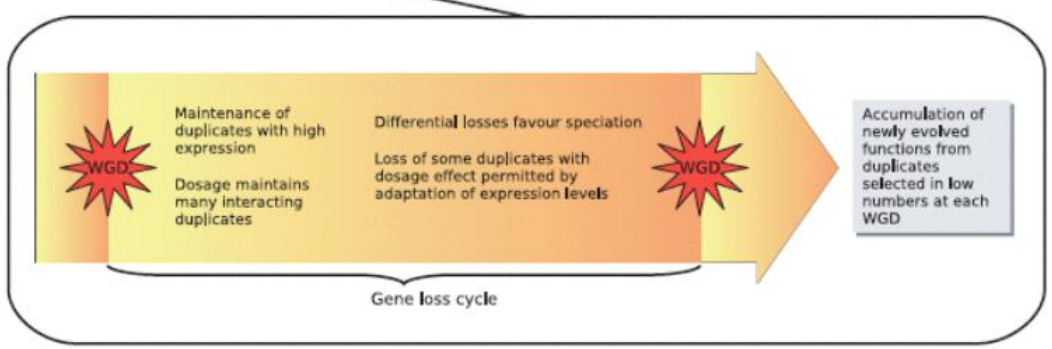


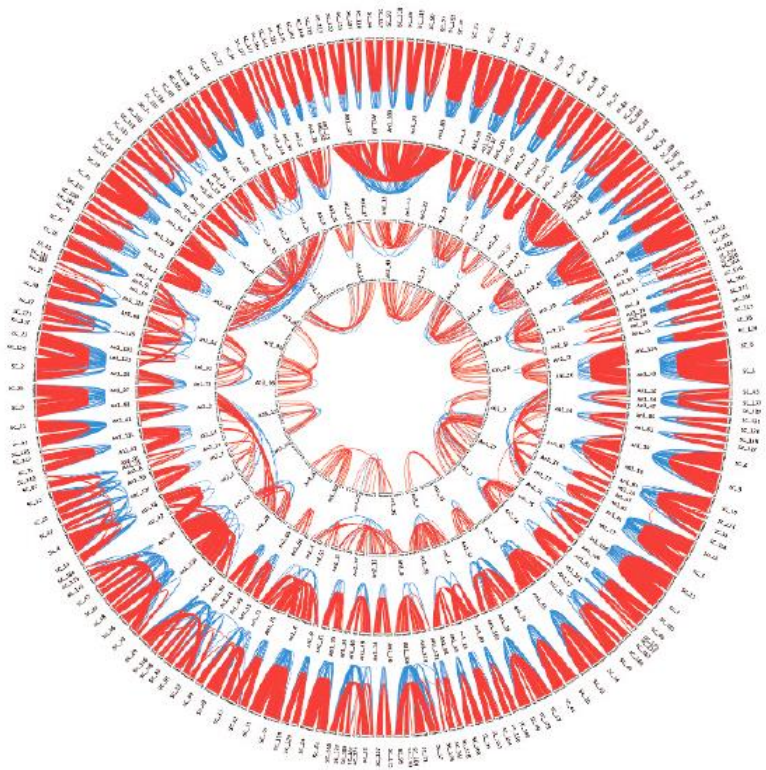
Comparison of two scaffolds originating from a common ancestor after a recent WGD

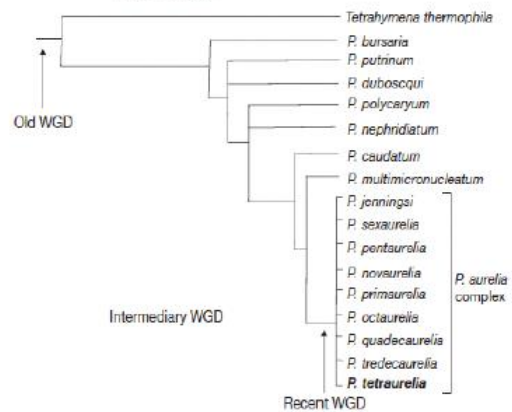
The macronuclear genome of *Paramecium tetraurelia*







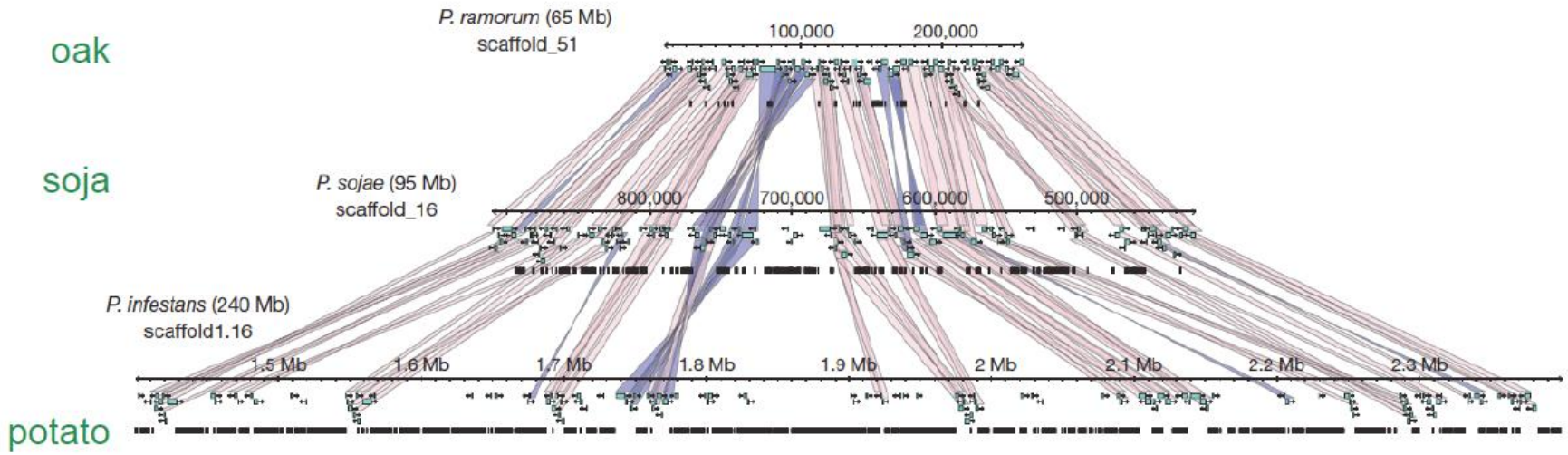




The genomes of *Phytophtora infestans (sojae* and *ramorum)*

Haas *et al.*, 2009 *Nature* **461**: 393-398

	P. infestans	*P. sojae*	*P. ramorum*
Genome size (Mb)	240	95	65
Scaffolds	4921	1810	2576
Repeat (%)	*74*	*39*	*28*
Protein-coding genes	17797	16988	14451



Conserved syntenic blocks containing most common genes and few repeated DNA are separated by regions of repeated DNA with low gene density and no conservation of gene order --> **dynamic genomes**

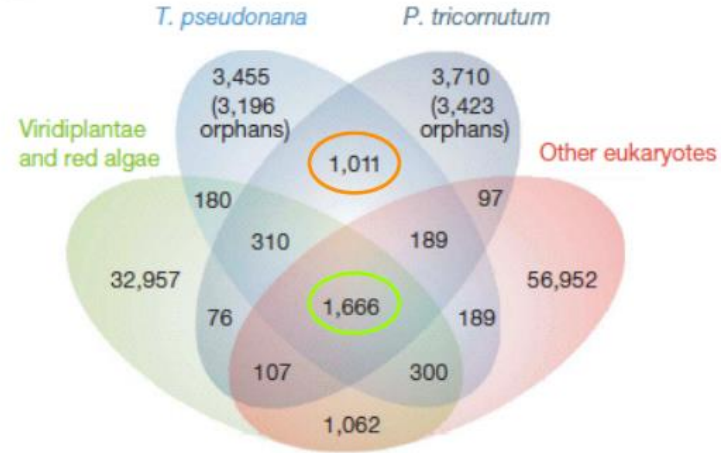
Rapidly evolving effector genes in non-conserved regions (modular secreted proteins, major types: RXLR and Crinkler, targetted to plant cells and responsible for necrosis, mostly species-specific)
Effector gene family expansion in *P. infestans* associated with numerous mobile elements (helitron)

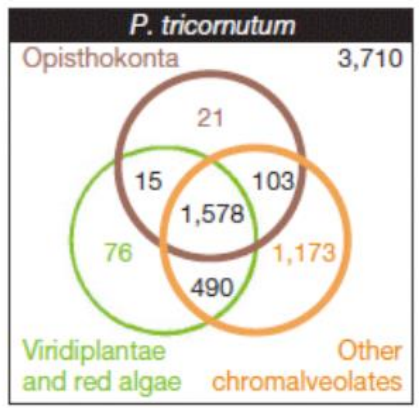
The genome of *Phaeodactylum tricornutum*

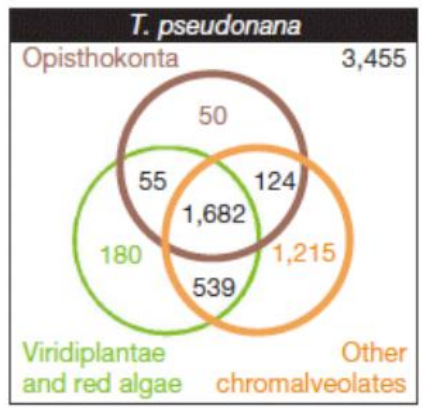
Bowler *et al.*, 2008, *Nature* **456**: 239-244

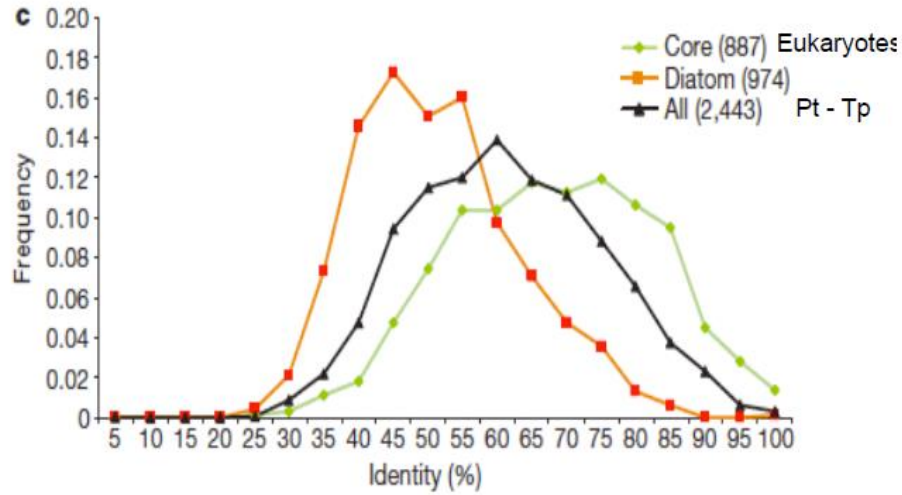
	pennate diatoms *P. tricornatum*	centric diatoms *Thalassiosira pseudonana*
Genome size (Mb)	27.4	32.4
Protein-coding genes	10402	11776
Spliceosomal introns	8169	17880

Numerous genes of bacterial origin involved in carbon and nitrogen utilization (xylanase, glucanase, prismane, carbon-nitrogen hydrolase, amidohydrolase), urea cycle (carbamoyl transferase, carabamate kinase, ornithine cyclodeaminase), cell wall silicification (S-adenosylmethionine-dependent decarboxylases and methyltransferases).

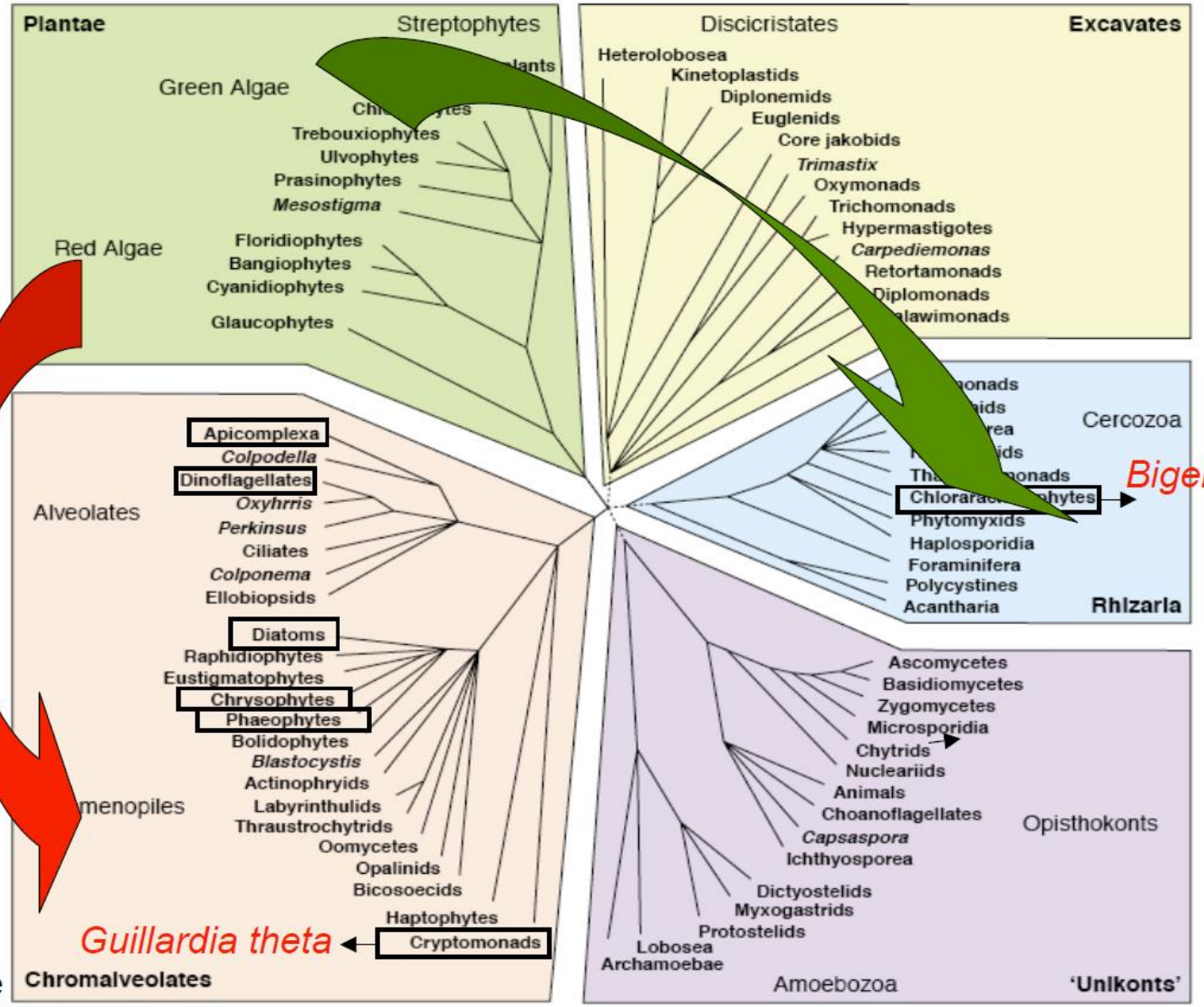








Four membrane plastids and nucleomorphs (double endosymbiosis)



red algae endosymbionte

Bigelowiella natans
green algae endosym-bionte

Nucleomorph genomes

Guillardia theta (Cryptomonads)

Douglas *et al.*, 2001 *Nature* **410**: 1091-1096

3 chromosomes (purified by PFGE) : **196, 181, 174 kb**
inverted repeats at chromosome ends (rDNA, ubiquitin-conjugating enzyme gene)

487 protein-coding genes
47 genes for non-coding RNAs (rRNAs, tRNAs, snRNAs, snoRNAs)
17 spliceosomal introns (42 - 52 nt)

compact genome (very short intergenic regions, partially overlapping genes)

Bigelowiella natans (Chlorarachiophytes)

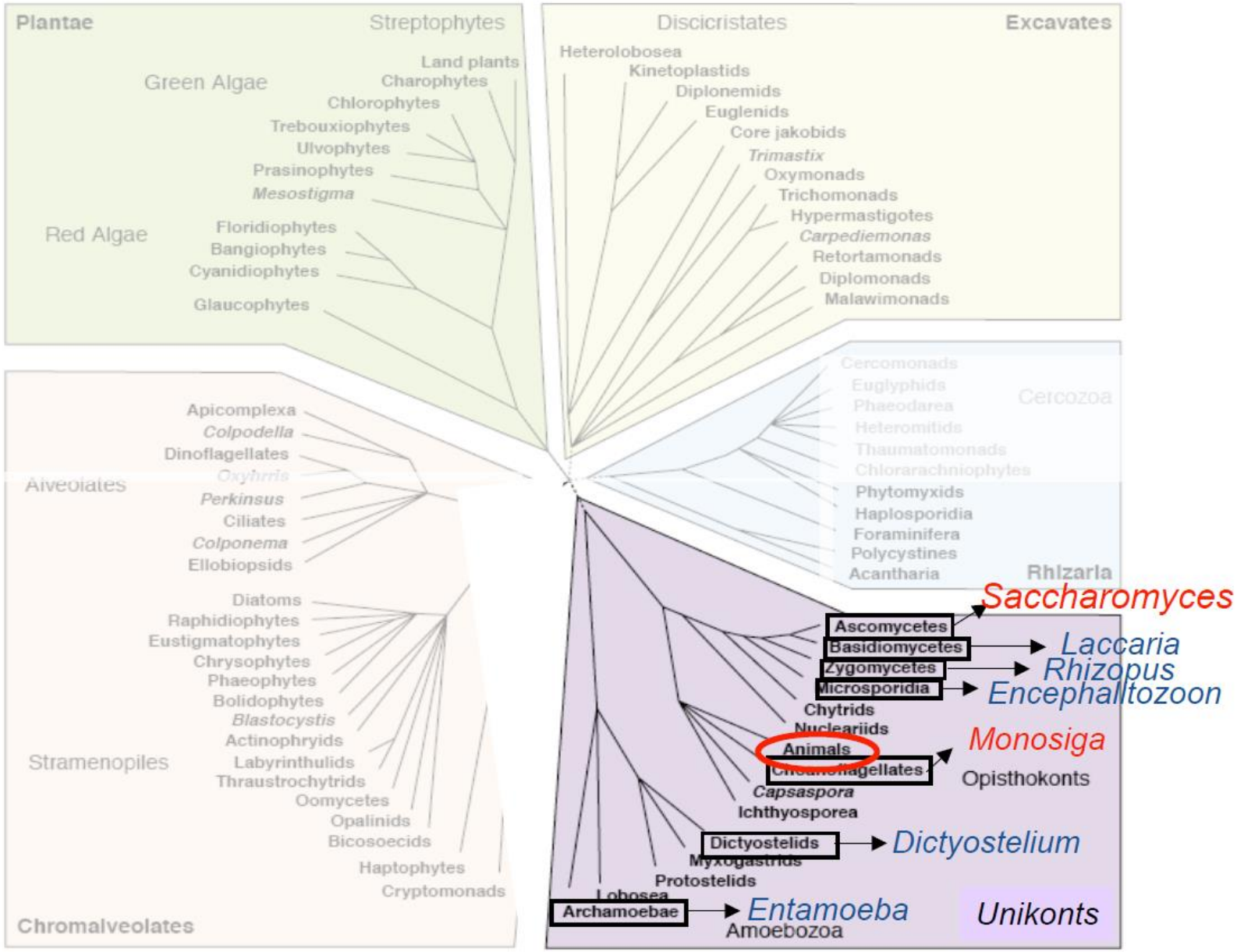
Gilson *et al.*, 2006 *PNAS* **103**: 9566-9571

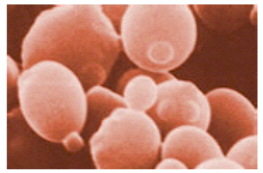
3 chromosomes (purified by PFGE): **141, 134, 98 kb**
inverted repeats at chromosome ends (rDNA, DnaK pseudogenes)

284 protein-coding genes (17 genes encoding plastid proteins)
42 genes for non-coding RNAs (rRNAs, tRNAs, snRNAs)

852 « pigmy » spliceosoaml introns (18 - 21 nt), splicing machinery

Genomes of unicellular eukaryotes: unikonts

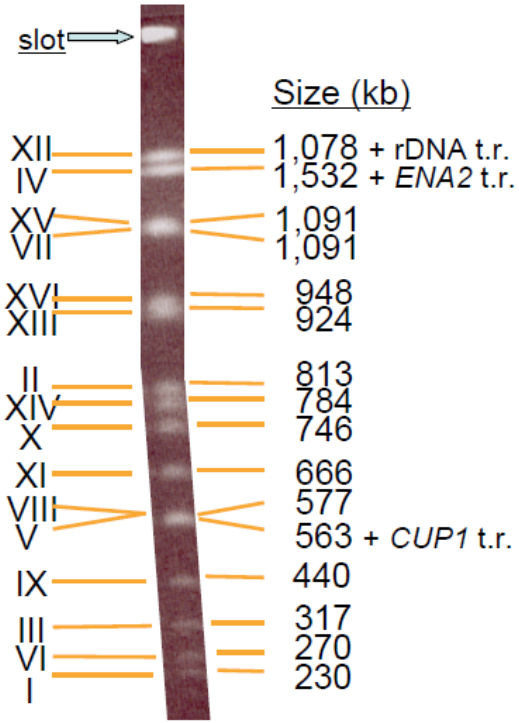




The exquisite beauty of a eukaryotic genome: *Saccharomyces cerevisiae*

1996

Nucleus: 16 chromosomes



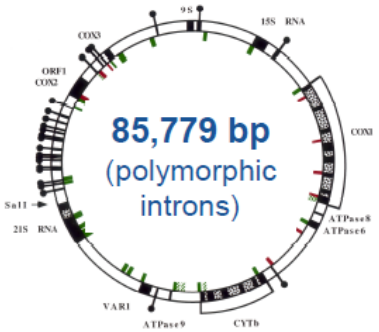
TOTAL 12,071 kb
+ tandem repeats

rDNA 9 kb repeat units
70-120 tandem copies
Total **0.6 - 1.1 Mb**

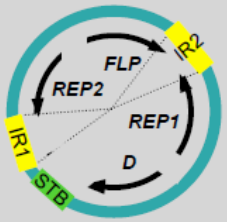
Dispersed repeated sequences

tRNA genes (identical)
Ty elements and LTR (similar)

Mitochondria



ca. 30 copies
Total **2.6 Mb**



two-micron **plasmid**
6,318 bp
ca. 50 copies
Total **0.3 Mb**

RNA "virus" 3 - 5 kb
100-1000 copies
Total **0.3 - 5 Mb**

	Total Noyau	Total Mitoch.
Chromosomes	**16**	**1**
Protein-coding genes		
active CDS	**5 769**	**7**
pseudogenes	77	-
RNA-coding genes		
transfer RNAs	275	23
sno RNAs	77	-
sn RNAs	6	-
ribosomal RNAs	3	2
Other RNAs	> 4	1
Mobile genetic elements		
complete (actifve)	52	-
incomplete (traces)	220	-
Introns		
spliceosomal	273	-
group I and group II	-	1 - 10

Yeast genomes are found in a variety of forms

Haploids, Aneuploids
Homozygous diploids, Heterozygous diploids
Interspecific hybrids, Partial hybrids, mosaics

A resource for functional data in eukaryotes
> 85 % of genes functionally characterized

Genome evolution in yeasts

Bernard Dujon[1], David Sherman[5,6], Gilles Fischer[1], Pascal Durrens[6,7], Serge Casaregola[8], Ingrid Lafontaine[1], Jacky de Montigny[9], Christian Marck[10], Cécile Neuvéglise[8], Emmanuel Talla[1], Nicolas Goffard[6], Lionel Frangeul[2], Michel Aigle[7], Véronique Anthouard[11], Anna Babour[8], Valérie Barbe[11], Stéphanie Barnay[8], Sylvie Blanchin[8], Jean-Marie Beckerich[8], Emmanuelle Beyne[5,6], Claudine Bleykasten[9], Anita Boisramé[8], Jeanne Boyer[1], Laurence Cattolico[11], Fabrice Confanioleri[12], Antoine de Daruvar[6], Laurence Despons[9], Emmanuelle Fabre[1], Cécile Fairhead[1], Hélène Ferry-Dumazet[6], Alexis Groppi[6], Florence Hantraye[3], Christophe Hennequin[1], Nicolas Jauniaux[9], Philippe Joyet[8], Rym Kachouri[13], Alix Kerrest[1], Romain Koszul[1], Marc Lemaire[14], Isabelle Lesur[5], Laurence Ma[2], Héloïse Muller[1], Jean-Marc Nicaud[8], Macha Nikolski[5], Sophie Oztas[11], Odile Ozier-Kalogeropoulos[1], Stefan Pellenz[1], Serge Potier[9], Guy-Franck Richard[1], Marie-Laure Straub[9], Audrey Suleau[8], Dominique Swennen[8], Fredj Tekaia[1], Micheline Wésolowski-Louvel[14], Eric Westhof[13], Bénédicte Wirth[9], Maria Zeniou-Meyer[9], Ivan Zivanovic[12], Monique Bolotin-Fukuhara[12], Agnès Thierry[1], Christiane Bouchier[2], Bernard Caudron[4], Claude Scarpelli[11], Claude Gaillardin[8], Jean Weissenbach[11], Patrick Wincker[11] & Jean-Luc Souciet[9]

[1]*Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR 927 Université Pierre et Marie Curie),* [2]*Plate-forme génomique, Pasteur Génopole Ile-de-France,* [3]*Unité de Génétique des interactions macromoléculaires (URA 2171 CNRS), and* [4]*Groupe Logiciels et Banques de données, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*
[5]*Laboratoire Bordelais de Recherche en Informatique (LaBRI, UMR 5800 CNRS), 351 cours de la Libération, 33405 Talence Cedex, France*
[6]*Centre de Bioinformatique de Bordeaux, Université Victor Ségalen (Bordeaux 2), 146 rue Léo Saignat, 33076 Bordeaux Cedex, France*
[7]*Institut de Biochimie et Génétique Cellulaires (UMR 5095 CNRS), Université Victor Ségalen (Bordeaux 2), 1 rue Camille Saint-Säens, 33077 Bordeaux Cedex, France*
[8]*Collection de Levures d'Intérêt Biotechnologique et Laboratoire de Génétique Moléculaire et Cellulaire (UMR 216 INRA and URA 1925 CNRS), INA-PG, PO Box 01, 78850 Thiverval-Grignon, France*
[9]*Laboratoire de Dynamique, Evolution et Expression des Génomes de Microorganismes (FRE 2326 CNRS), Université Louis Pasteur, 28 rue Goethe, 67000 Strasbourg, France*
[10]*Service de Biochimie et de Génétique Moléculaire, CEA/Saclay, 91191 Gif-sur Yvette, France*
[11]*Génoscope (UMR 8030 CNRS), 2 rue Gaston Crémieux, 91057 Evry Cedex, France*
[12]*Institut de Génétique Moléculaire (UMR 8621 CNRS), Université de Paris Sud, Bâtiment 400, 91405 Orsay Cedex, France*
[13]*Modélisations et Simulations des Acides Nucléiques, IBMC (UPR 9002 CNRS), 15 rue René Descartes, 67000 Strasbourg, France*
[14]*Laboratoire de Génétique des Levures (UMR 5122 CNRS), Université Claude Bernard, Bâtiment Lwoff, 43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France*

Identifying the mechanisms of eukaryotic genome evolution by comparative genomics is often complicated by the multiplicity of events that have taken place throughout the history of individual lineages, leaving only distorted and superimposed traces in the genome of each living organism. The hemiascomycete yeasts, with their compact genomes, similar lifestyle and distinct sexual and physiological properties, provide a unique opportunity to explore such mechanisms. We present here the complete, assembled genome sequences of four yeast species, selected to represent a broad evolutionary range within a single eukaryotic phylum, that after analysis proved to be molecularly as diverse as the entire phylum of chordates. A total of approximately 24,200 novel genes were identified, the translation products of which were classified together with *Saccharomyces cerevisiae* proteins into about 4,700 families, forming the basis for interspecific comparisons. Analysis of chromosome maps and genome redundancies reveal that the different yeast lineages have evolved through a marked interplay between several distinct molecular mechanisms, including tandem gene repeat formation, segmental duplication, a massive genome duplication and extensive gene loss.

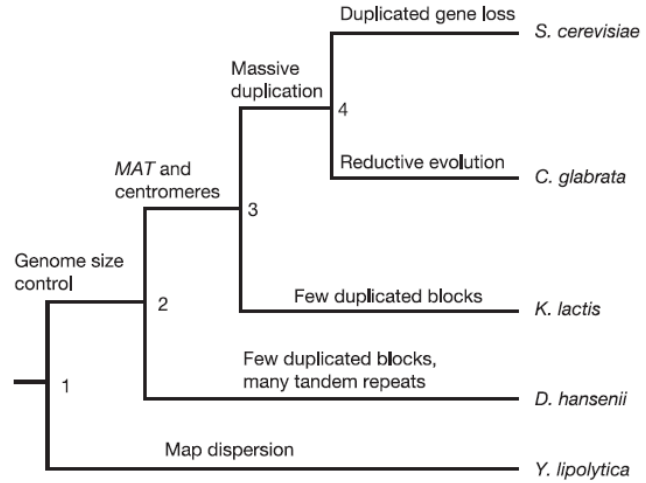


Figure 6 Major evolutionary events in the genomes of hemiascomycetes. Shown is a cartoon of the evolutionary history of the four sequenced yeasts, plus *S. cerevisiae*. The tree topology is based on 25S rDNA sequences (see Supplementary Fig. S17). The most conspicuous evolutionary signatures are summarized on each branch. The tendency for map dispersion of *Y. lipolytica* is visible at a variety of levels (see text). The other yeasts share an ability to duplicate genes in an ordered manner. An accidental whole-genome duplication event occurred in the common ancestor of *S. cerevisiae* and *C. glabrata*. It has been followed by extensive gene loss of paralogous copies.

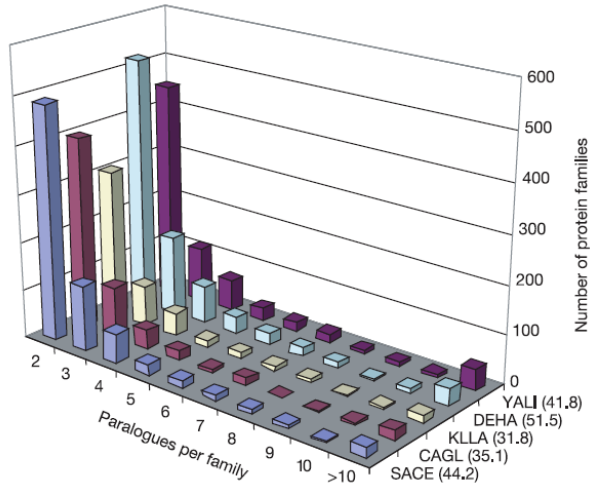


Figure 1 Overall genome redundancy as deduced from protein families. Shown for each yeast species is the total number of protein families distributed according to their size. A similar pattern (not shown) is obtained when considering only the universal protein families (sckdy pattern). The overall genome redundancy for each species, defined as the ratio (in per cent) of the number of CDS belonging to multigene families over the total number of CDS, is indicated in brackets next to the species abbreviation. SACE, *S. cerevisiae*; CAGL, *C. glabrata*; KLLA, *K. lactis*; DEHA, *D. hansenii*; YALI, *Y. lipolytica*.

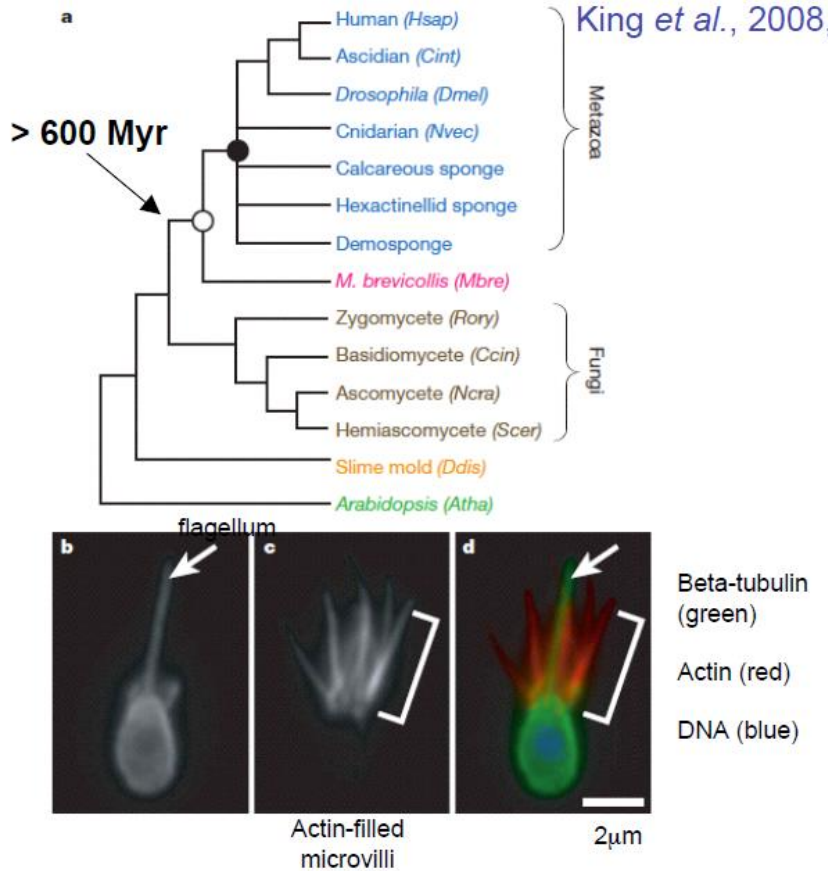
Table 2 **General characteristics of the yeast genomes and predicted proteomes**

Species	Genome size (Mb)	Average G+C content (%)	Total CDS	Total tRNA genes	Average gene density (%)	Average G+C in CDS (%)	Average CDS size (codons)	Median CDS size (codons)	Maximum CDS size (codons)
S. cerevisiae	12.1	38.3	5,807	274	70.3	39.6	485	398	4,911
C. glabrata	12.3	38.8	5,283	207	65.0	41.0	493	409	4,881
K. lactis	10.6	38.7	5,329	162	71.6	40.1	461	381	4,916
D. hansenii	12.2	36.3	6,906	205	79.2	37.5	389	307	4,190
Y. lipolytica	20.5	49.0	6,703	510	46.3	52.9	476	399	6,539

Figures are calculated from final chromosome sequences or scaffolds, after annotation. Genome sizes do not include rDNA. Average gene density represents the fraction of each genome occupied by the protein-coding genes (other genetic elements are not considered). Figures for *D. hansenii* are only tentative; figures for *S. cerevisiae* were recently recomputed from http://mips.gsf.de/genre/proj/yeast.

The genome of the choanoflagellate *Monosiga brevicollis*

King *et al.*, 2008, *Nature* **451**: 783-788



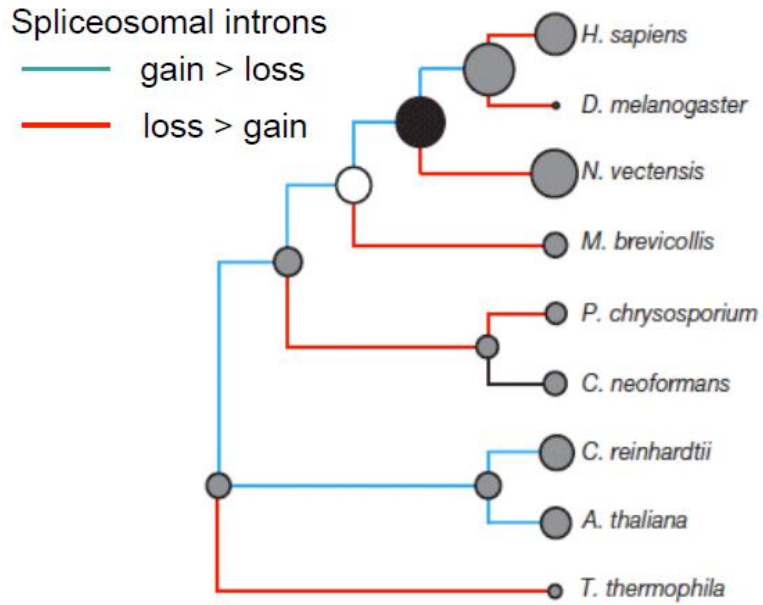


Table 1 | *M. brevicollis* genome properties in a phylogenetic context

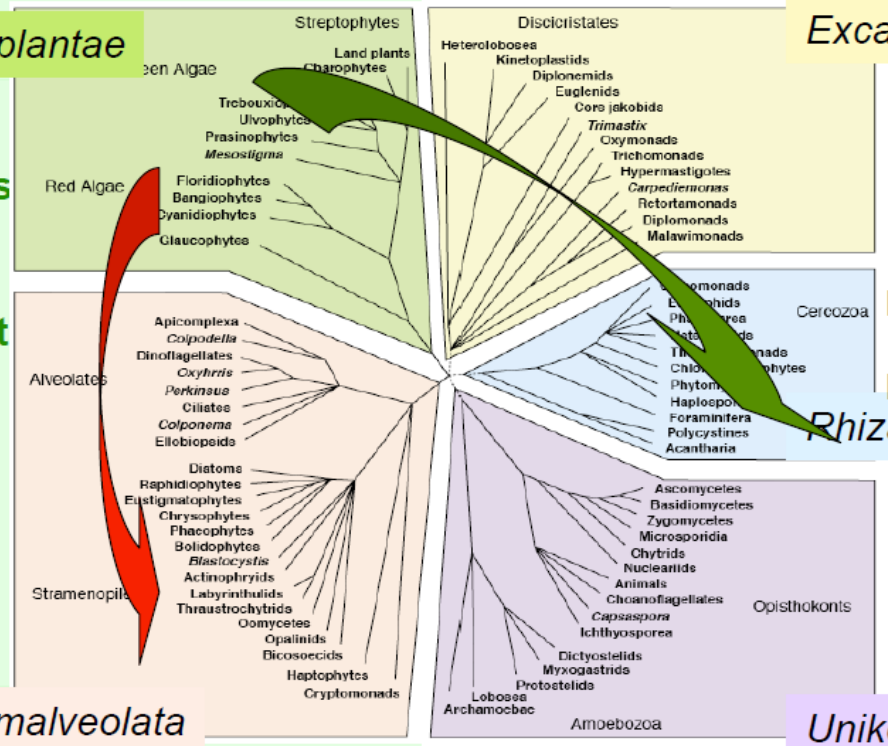
	Metazoa				Choanoflagellates	Fungi		*Dictyostelium*	Plants
	Hsap	*Cint*	*Dmel*	*Nvec*	*Mbre*	*Ccin*	*Ncra*	*Ddis*	*Atha*
Genome size (Mb)	2,900	160	180	357	42	38	39	34	125
Total number of genes	23,224	14,182	14,601	18,000	9,196	13,544	9,826	13,607	27,273
Mean gene size (bp)	27,000	4,585	5,247	6,264	3,004	1,679	1,528	1,756	2,287
Mean intron density (introns per gene)	7.7	6.8	4.9	5.8	6.6	4.4	1.8	1.9	4.4
Mean intron length (bp)	3,365	477	1,192	903	174	75	136	146	164
Gene density (kb per gene)	127.9	11.9	13.2	19.8	4.5	2.7	4.0	2.5	4.5

The world of unicellular eukaryotes from genomics

Small, compact genomes
Transposon clustering --> compositional heterogeneity
Numerous gene fusions
Horizontal gene acquisitions
Multicellularity

Primary endosymbiosis (chloroplasts)

followed by chloroplast loss



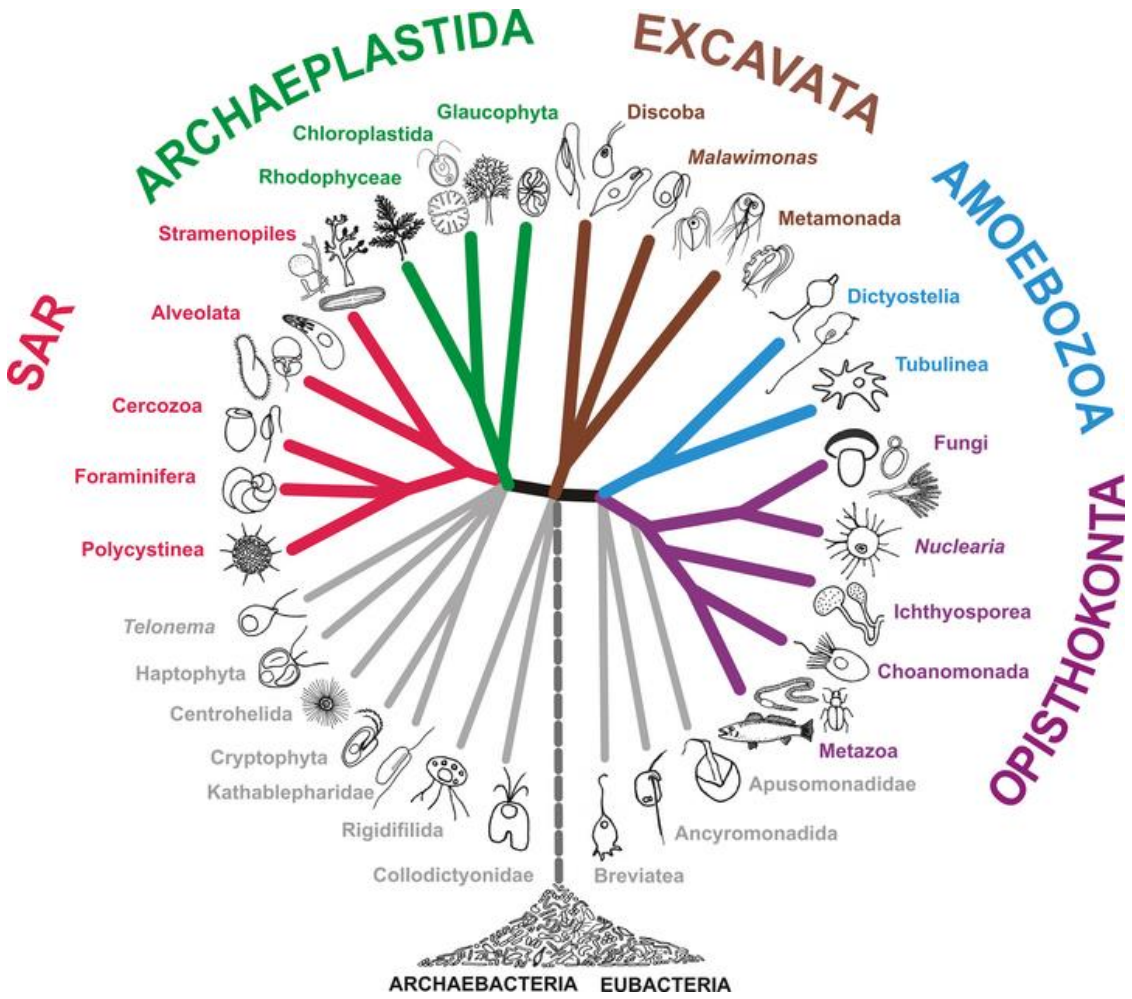
Small, compact genomes
Simplified molecular machinery
Regressive evolution
Horizontal gene acquisitions

Double endosymbionts (green algae)
Horizontal gene acquisitions
else ?

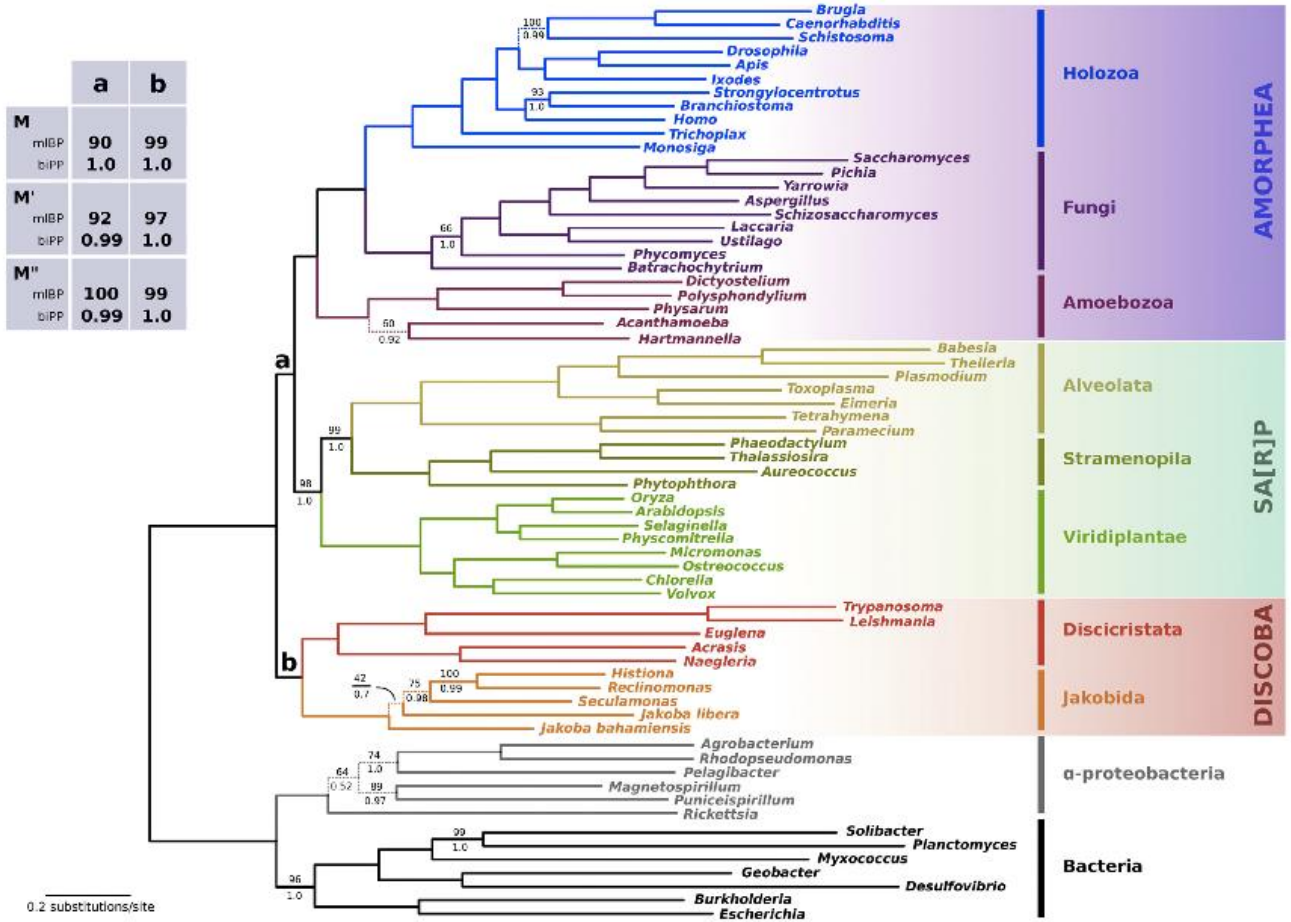
Whole-genome duplications and gene loss
Rapidly evolving dynamic genomes
Transposon clustering
Horizontal gene acquisitions
Double endosymbionts (red algae)

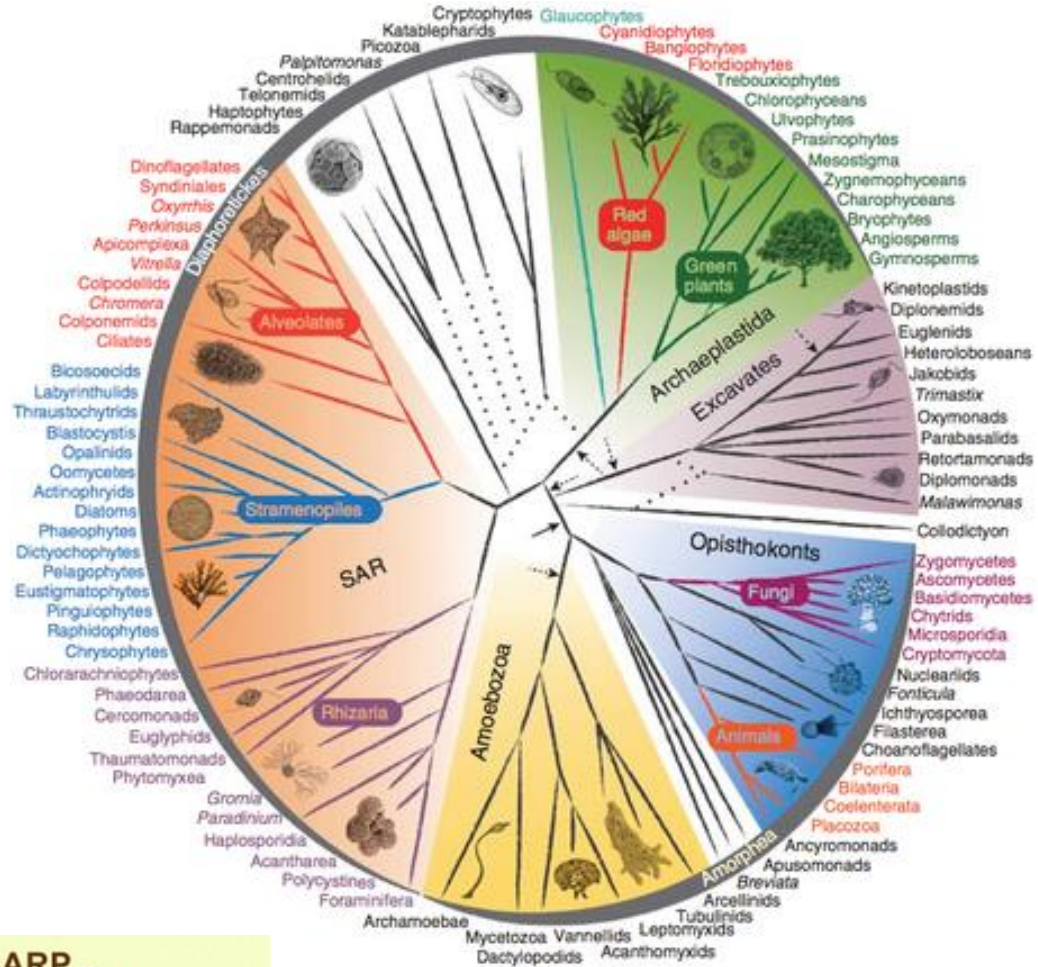
Intron gain and loss
Numerous gene fusions
Horizontal gene acquisitions
Multicellularity
Whole-genome duplications

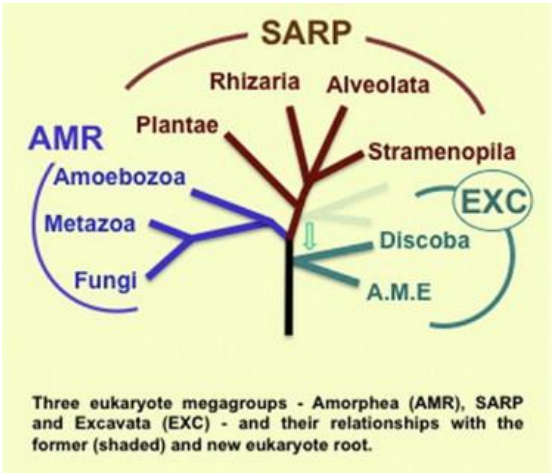
Filogenija in evolucijski odnosi eukariontov



Supergroup	Representatives: Groups	Representatives: Genera	Assessment: Nuclear	Assessment: Plastid
"Amoebozoa"	Lobose amoebae Slime molds Pelobionts	*Amoeba* *Dictyostelium* *Entamoeba*	+	
"Chromalveolata"	Ciliates Stramenopiles Apicomplexa	*Tetrahymena* *Phytophthora* *Plasmodium*	-	+
"Excavata"	Diplomonads Euglenozoa Parabasalids	*Giardia* *Trypanosoma* *Trichomonas*	-	
"Opisthokonta"	Animals Fungi Choanoflagellates	*Drosophila* *Encephalitozoon* *Monosiga*	+ + +	
"Plantae"	Green algae Red algae Glaucophytes	*Arabidopsis* *Porphyra* *Cyanophora*	+	+ + +
"Rhizaria"	Cercomonads Foraminifera Euglyphids	*Cercomonas* *Allogromia* *Paulinella*	+	







Three eukaryote megagroups - Amorphea (AMR), SARP and Excavata (EXC) - and their relationships with the former (shaded) and new eukaryote root.

He et al. An alternative root for the eukaryote tree of life. Curr Biol. 2014 Feb 17;24(4):465-470.

Rastlinski genomi

-Genomi rastlin, ki so pomembni za prehrano (žitarice: riž, koruza itd),
-orjaški genomi golosemenk (iglavcev)
-Genomi prvotnih kopenski rastlin,
-Genomi zelenih in rdečih alg.

Glivni genomi

-Genomi kvasovk,
-Genomi patogenih gliv,
-Genomi biotehnološko pomembnih gliv,
-Genomi najstarejših gliv so podobni živalskim genomom.

Živalski genomi

-Genom človeka,
-Genomi arhaičnih ljudi (Neandertalci in Denizovanci),
-Genomi primatov,
-Genomi različnih sesalcev,
-Genomi vretenčarjev (od piškurja do plazilcev/ptičev),
-Genomi nevretenčarjev,
-Najstarejši živalski genomi (spužve, ožigalkarji in rebrenjače).

Genomi „primitivnih“ eukariontov

-Genomi parazitov in patogenov,
-Genomi prostoživečih organizmov,
-Kaj je vseboval genom najstarejših/prvotnih eukariontov.

Rastlinski genomi

The B73 Maize Genome: Complexity, Diversity, and Dynamics

We report an improved draft nucleotide sequence of the 2.3-gigabase genome of maize, an important crop plant and model for biological research. Over 32,000 genes were predicted, of which 99.8% were placed on reference chromosomes. Nearly 85% of the genome is composed of hundreds of families of transposable elements, dispersed nonuniformly across the genome. These were responsible for the capture and amplification of numerous gene fragments and affect the composition, sizes, and positions of centromeres. We also report on the correlation of methylation-poor regions with *Mu* transposon insertions and recombination, and copy number variants with insertions and/or deletions, as well as how uneven gene losses between duplicated regions were involved in returning an ancient allotetraploid to a genetically diploid state. These analyses inform and set the stage for further investigations to improve our understanding of the domestication and agricultural improvements of maize.

Analysis of the bread wheat genome using whole-genome shotgun sequencing

Bread wheat (***Triticum aestivum***) is a globally important crop, accounting for 20 per cent of the calories consumed by humans. Major efforts are underway worldwide to increase wheat production by extending genetic diversity and analysing key traits, and genomic resources can accelerate progress. But so far the very large size and polyploid complexity of the bread wheat genome have been substantial barriers to genome analysis. Here we report the sequencing of its large, 17-gigabase-pair, hexaploid genome using 454 pyrosequencing, and comparison of this with the sequences of diploid ancestral and progenitor genomes. We identified between 94,000 and 96,000 genes, and assigned two-thirds to the three component genomes (A, B and D) of hexaploid wheat. High-resolution synteny maps identified many small disruptions to conserved gene order. We show that the hexaploid genome is highly dynamic, with significant loss of gene family members on polyploidization and domestication, and an abundance of gene fragments. Several classes of genes involved in energy harvesting, metabolism and growth are among expanded gene families that could be associated with crop productivity. Our analyses, coupled with the identification of extensive genetic variation, provide a resource for accelerating gene discovery and improving this major crop.

The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important syntenic relationships with the other cereal species and is a model plant for the grasses. Here we present a map-based, finished quality sequence that covers 95% of the 389 Mb genome, including virtually all of the euchromatin and two complete centromeres. A total of 37,544 non-transposable-element-related protein-coding genes were identified, of which 71% had a putative homologue in ***Arabidopsis***. In a reciprocal analysis, 90% of the ***Arabidopsis*** proteins had a putative homologue in the predicted rice proteome. Twenty-nine per cent of the 37,544 predicted genes appear in clustered gene families. The number and classes of transposable elements found in the rice genome are consistent with the expansion of syntenic regions in the maize and sorghum genomes. We find evidence for widespread and recurrent gene transfer from the organelles to the nuclear chromosomes. The map-based sequence has proven useful for the identification of genes underlying agronomic traits. The additional single-nucleotide polymorphisms and simple sequence repeats identified in our study should accelerate improvements in rice production.

Genome sequence and analysis of the tuber crop potato

The Potato Genome Sequencing Consortium*

Potato (***Solanum tuberosum*** L.) is the world's most important non-grain food crop and is central to global food security. It is clonally propagated, highly heterozygous, autotetraploid, and suffers acute inbreeding depression. Here we use a homozygous doubled-monoploid potato clone to sequence and assemble 86% of the 844-megabase genome. We predict 39,031 protein-coding genes and present evidence for at least two genome duplication events indicative of a palaeopolyploid origin. As the first genome sequence of an asterid, the potato genome reveals 2,642 genes specific to this large angiosperm clade. We also sequenced a heterozygous diploid clone and show that gene presence/absence variants and other potentially deleterious mutations occur frequently and are a likely cause of inbreeding depression. Gene family expansion, tissue-specific expression and recruitment of genes to new pathways contributed to the evolution of tuber development. The potato genome sequence provides a platform for genetic improvement of this vital crop.

The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla

The French–Italian Public Consortium for Grapevine Genome Characterization*

The analysis of the first plant genomes provided unexpected evidence for genome duplication events in species that had previously been considered as true diploids on the basis of their genetics[1–3]. These polyploidization events may have had important consequences in plant evolution, in particular for species radiation and adaptation and for the modulation of functional capacities[4–10]. Here we report a high-quality draft of the genome sequence of grapevine (*Vitis vinifera*) obtained from a highly homozygous genotype. The draft sequence of the grapevine genome is the fourth one produced so far for flowering plants, the second for a woody species and the first for a fruit crop (cultivated for both fruit and beverage). Grapevine was selected because of its important place in the cultural heritage of humanity beginning during the Neolithic period[11]. Several large expansions of gene families with roles in aromatic features are observed. The grapevine genome has not undergone recent genome duplication, thus enabling the discovery of ancestral traits and features of the genetic organization of flowering plants. This analysis reveals the contribution of three ancestral genomes to the grapevine haploid content. This ancestral arrangement is common to many dicotyledonous plants but is absent from the genome of rice, which is a monocotyledon. Furthermore, we explain the chronology of previously described whole-genome duplication events in the evolution of flowering plants.

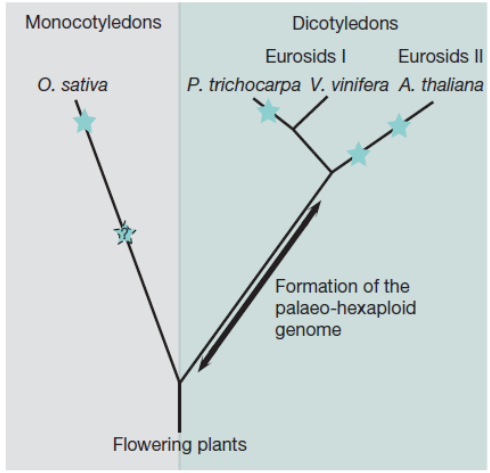


Figure 3 | Positions of the polyploidization events in the evolution of plants with a sequenced genome. Each star indicates a WGD (tetraploidization) event on that branch. The question mark indicates that ancient events are visible in the rice genome that would require other monocotyledon genome sequences to be resolved. The formation of the palaeo-hexaploid ancestral genome occurred after divergence from monocotyledons and before the radiation of the Eurosids.

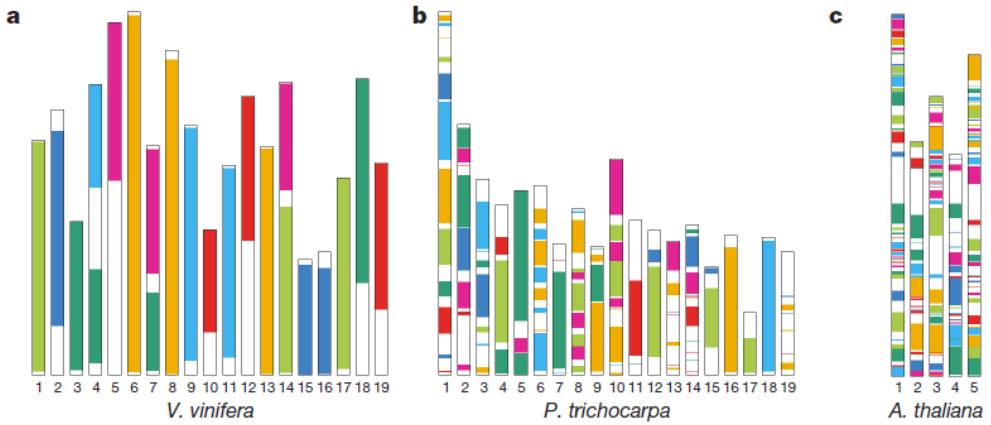


Figure 2 | Schematic representation of paralogous regions derived from the three ancestral genomes in the karyotypes of *V. vinifera*, *P. trichocarpa* and *A. thaliana*. Each colour corresponds to a syntenic region between the three ancestral genomes that were defined by their occurrence as linked clusters in grapevine, independently of intrachromosomal rearrangements. The *V. vinifera* genome (**a**) is by far the closest to the ancestral arrangement, whereas that of *Arabidopsis* (**c**) is thoroughly rearranged, and *P. trichocarpa* (**b**) presents an intermediate situation. The seven colours probably correspond to linkage groups at the time of the palaeo-hexaploid ancestor.

The Norway spruce genome sequence and conifer genome evolution

Lists of authors and their affiliations appear at the end of the paper

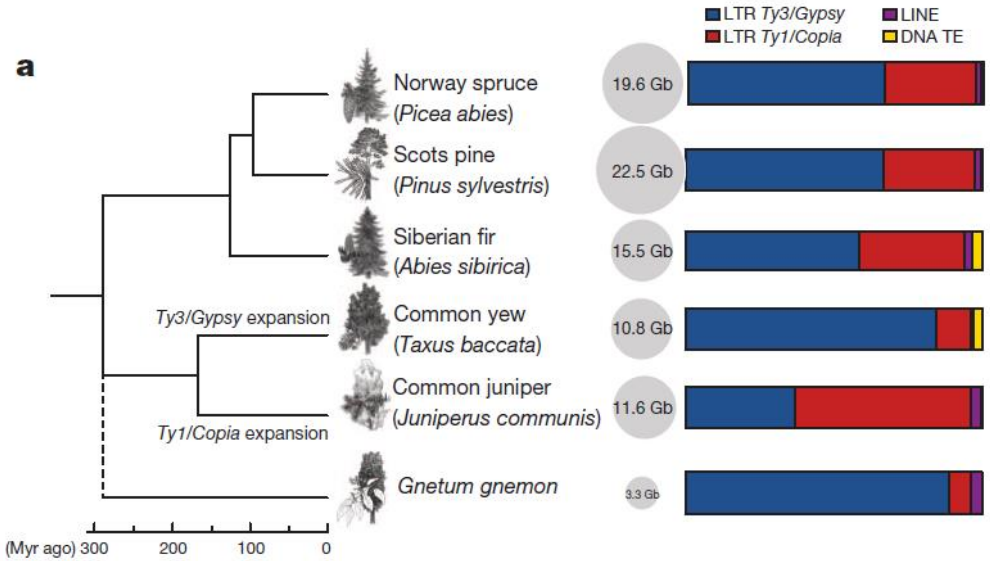
Conifers have dominated forests for more than 200 million years and are of huge ecological and economic importance. Here we present the draft assembly of the 20-gigabase genome of Norway spruce (*Picea abies*), the first available for any gymnosperm. The number of well-supported genes (28,354) is similar to the >100 times smaller genome of *Arabidopsis thaliana*, and there is no evidence of a recent whole-genome duplication in the gymnosperm lineage. Instead, the large genome size seems to result from the slow and steady accumulation of a diverse set of long-terminal repeat transposable elements, possibly owing to the lack of an efficient elimination mechanism. Comparative sequencing of *Pinus sylvestris*, *Abies sibirica*, *Juniperus communis*, *Taxus baccata* and *Gnetum gnemon* reveals that the transposable element diversity is shared among extant conifers. Expression of 24-nucleotide small RNAs, previously implicated in transposable element silencing, is tissue-specific and much lower than in other plants. We further identify numerous long (>10,000 base pairs) introns, gene-like fragments, uncharacterized long non-coding RNAs and short RNAs. This opens up new genomic avenues for conifer forestry and breeding.

Table 1 | Characteristics of the *P. abies* genome

Genome	
Size (1*n*)	19.6 Gb
Karyotype	2*n* = 24
GC content	37.9%
High-copy repeat content*	
LTR_*Gypsy/Copia*/unclassified	35%/16%/7%
LINE	1%
DNA transposable element	1%
Unclassified	10%
Genes and gene-like fragments†	2.4%
Assembly	
Size in scaffolds >200 bp/>10 kb	12 Gb/4.3 Gb
N50/NG50	4,869 bp/721 bp
Annotation	
High confidence gene set	28,354
Genes with >5-kb introns	8.4%
Avg. exon/intron size	312 bp/1,017 bp
Avg. gene density	1 gene in 705 kb
Transposable element genes	284,587
Non-coding loci	
lncRNA (unique/conserved)	13,031/9,686
miRNA (*de novo* predicted)	2,719

*Inferred from unassembled reads. †Including pseudogenes, excluding transposable elements.

Norway spruce genome sequenced: Largest ever to be mapped



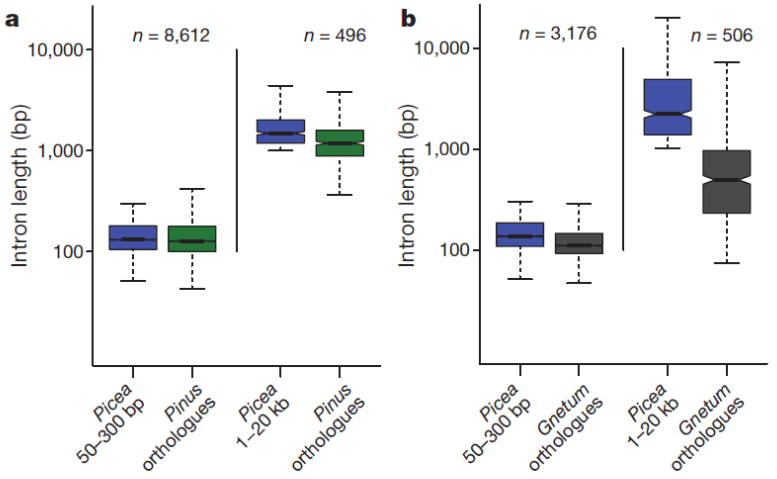


Figure 3 | Intron sizes are conserved among gymnosperms. a, b, Intron size comparisons between *P. abies*, *P. sylvestris* (**a**) and *G. gnemon* (**b**), respectively. Orthologues of introns that were categorised as short (50–300 bp) or long (1–20 kb) in *P. abies* were identified in *P. sylvestris* and *G. gnemon*, and the corresponding intron size was scored.

Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies

Abstract

Background: The size and complexity of conifer genomes has, until now, prevented full genome sequencing and assembly. The large research community and economic importance of loblolly pine, *Pinus taeda* L., made it an early candidate for reference sequence determination.

Results: We develop a novel strategy to sequence the genome of loblolly pine that combines unique aspects of pine reproductive biology and genome assembly methodology. We use a whole genome shotgun approach relying primarily on next generation sequence generated from a single haploid seed megagametophyte from a loblolly pine tree, 20-1010, that has been used in industrial forest tree breeding. The resulting sequence and assembly was used to generate a draft genome spanning 23.2 Gbp and containing 20.1 Gbp with an N50 scaffold size of 66.9 kbp, making it a significant improvement over available conifer genomes. The long scaffold lengths allow the annotation of 50,172 gene models with intron lengths averaging over 2.7 kbp and sometimes exceeding 100 kbp in length. Analysis of orthologous gene sets identifies gene families that may be unique to conifers. We further characterize and expand the existing repeat library based on the *de novo* analysis of the repetitive content, estimated to encompass 82% of the genome.

Conclusions: In addition to its value as a resource for researchers and breeders, the loblolly pine genome sequence and assembly reported here demonstrates a novel approach to sequencing the large and complex genomes of this important group of plants that can now be widely applied.

Table 2 Comparison of gene metrics among sequenced plant genomes

	Pinus taeda	*Picea abies* [8]	*Arabidopsis thaliana* [21]	*Populus trichocarpa* [21]	*Vitis vinifera* [21]	*Amborella trichopoda* [22]
Genome size (assembled) (Mbp)	20,148	12,019[a]	135	423	487	706
Chromosomes	12	12	5	19	19	13
G + C content (%)	38.2	37.9	35.0	33.3	36.2	35.5
TE content (%)	79	70	15.3	42	41.4	N/A
Number of genes[b]	50,172	58,587[c]	27,160	36,393	25,663	25,347
Average CDS length (bps)	965	723	1102	1143	1095	969
Average intron length (bps)	2,741	1,020	182	366	933	1,538
Maximum intron length (bps)	318,524	68,269	10,234	4,698	38,166	175,748

[a]Estimated genome size is 19.6 Gbp.
[b]Number of full-length genes >150 bp in length and validated through current annotations.
[c]High and medium confidence genes from the Congenie project [8].

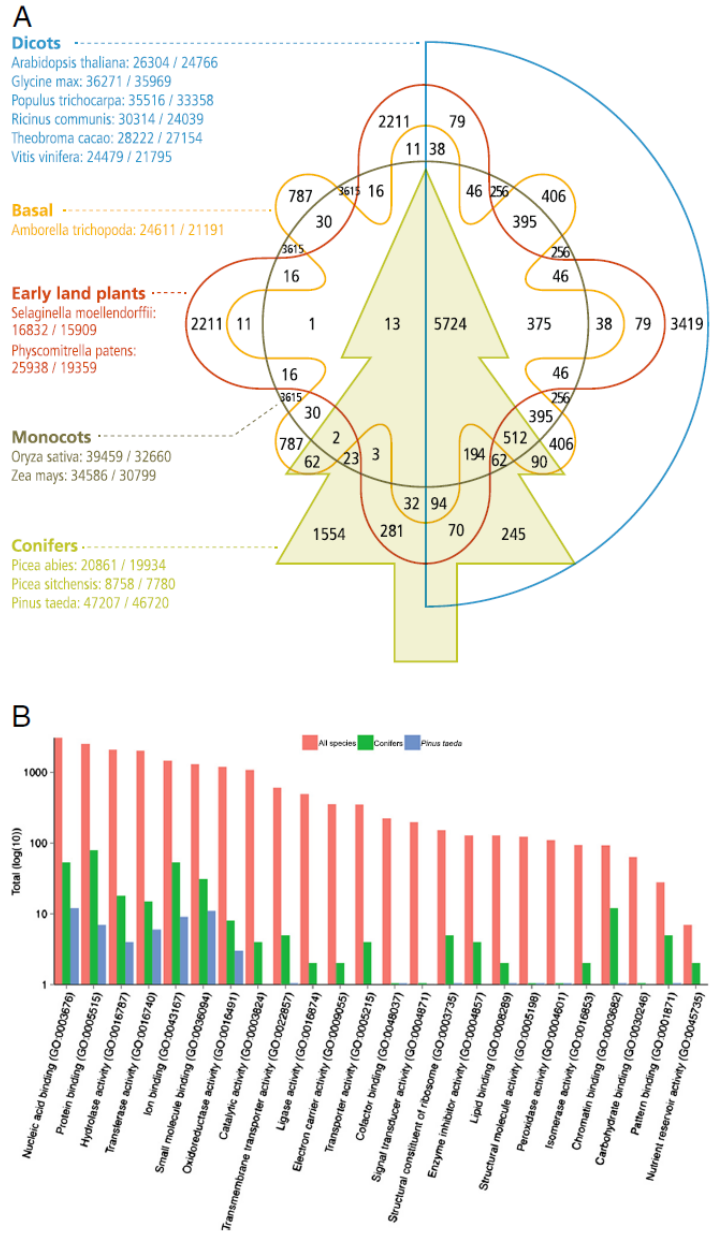
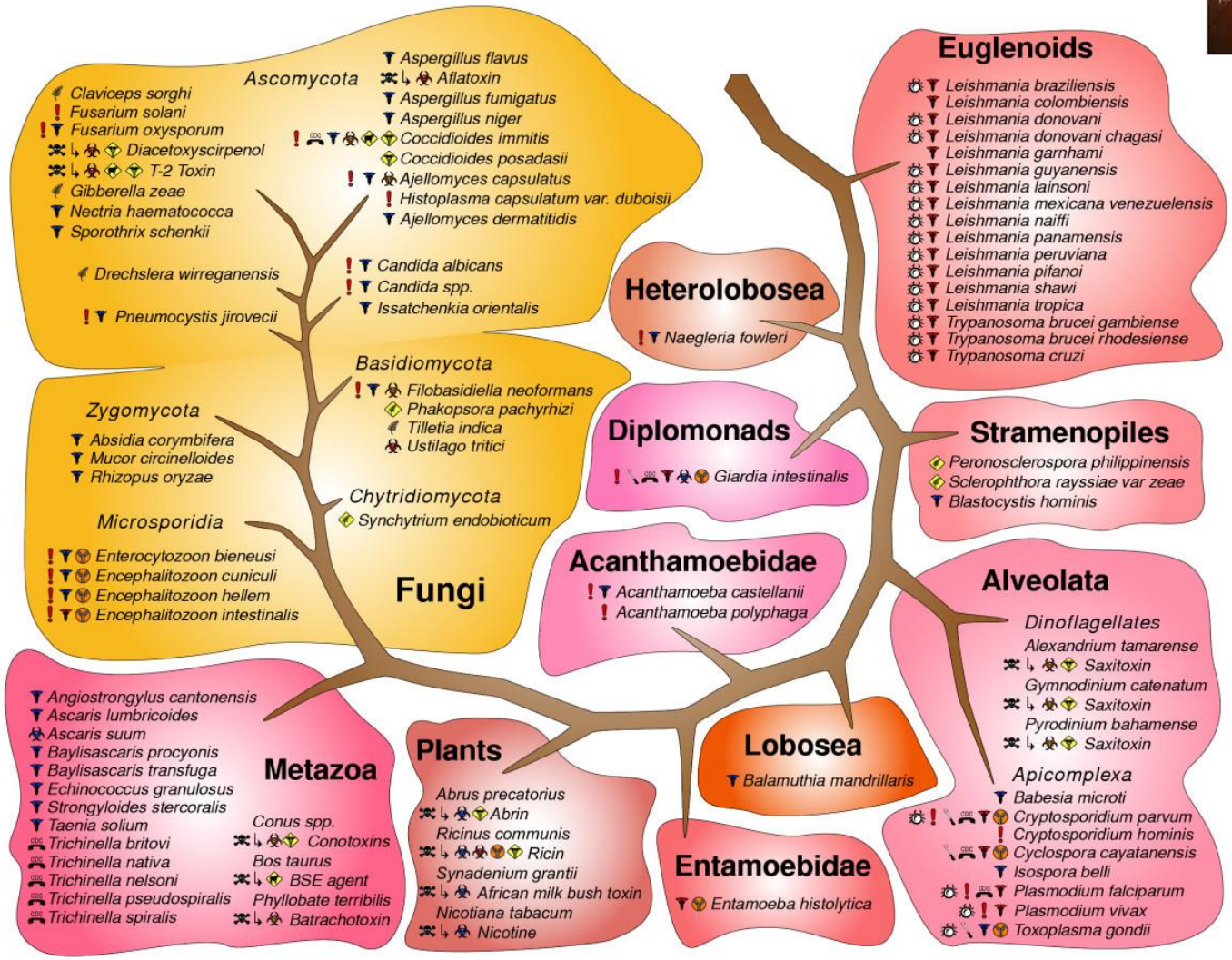


Figure 2 Unique gene families and Gene Ontology term assignments. (A) Identification of orthologous groups of genes for 14 species split into five categories: conifers (*Picea abies*, *Picea sitchensis*, and *Pinus taeda*), monocots (*Oryza sativa* and *Zea mays*), dicots (*Arabidopsis thaliana*, *Glycine max*, *Populus trichocarpa*, *Ricinus communis*, *Theobroma cacao*, and *Vitis vinifera*), early land plants (*Selaginella moellendorffii* and *Physcomitrella patens*), and a basal angiosperm (*Amborella trichopoda*). Here, we depict the number of clusters in common between the biological categories in the intersections. The total number of sequences for each species is provided under the name (total number of sequences/total number of clustered sequences). **(B)** Gene ontology molecular function term assignments by family for all species (red), conifers (green), and *Pinus taeda* exclusively (blue).

Genomi parazitov in patogenov



Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*

Aspergillus fumigatus is exceptional among microorganisms in being both a primary and opportunistic pathogen as well as a major allergen[1–3]. Its conidia production is prolific, and so human respiratory tract exposure is almost constant[4]. *A. fumigatus* is isolated from human habitats[5] and vegetable compost heaps[6,7]. In immunocompromised individuals, the incidence of invasive infection can be as high as 50% and the mortality rate is often about 50% (ref. 2). The interaction of *A. fumigatus* and other airborne fungi with the immune system is increasingly linked to severe asthma and sinusitis[8]. Although the burden of invasive disease caused by *A. fumigatus* is substantial, the basic biology of the organism is mostly obscure. Here we show the complete 29.4-megabase genome sequence of the clinical isolate Af293, which consists of eight chromosomes containing 9,926 predicted genes. Microarray analysis revealed temperature-dependent expression of distinct sets of genes, as well as 700 *A. fumigatus* genes not present or significantly diverged in the closely related sexual species *Neosartorya fischeri*, many of which may have roles in the pathogenicity phenotype. The Af293 genome sequence provides an unparalleled resource for the future understanding of this remarkable fungus.

Table 1 | Properties of the *Aspergillus fumigatus* Af293 genome

Genome	Value
Nuclear genome	
General information	
Size (Mb)	29.4
G+C content (%)	49.9
Gene number	9,926
Mean gene length (bp)	1,431
Per cent coding	50.1
Per cent genes with introns	77.0
Genes of unknown function	3,288
Exons	
Mean number per gene	2.8
Mean length (bp)	516
G+C content (%)	54.0
Introns	
Mean number per gene	1.8
Mean length (bp)	112
G+G content (%)	46.3
Intergenic regions	
Mean length (bp)	1,226
G+C content (%)	46.0
RNA	
tRNA number	179
5S rRNA number	33
Mitochondrial genome	
Size (bp)	31,892
G+C content (%)	25.4
Gene number	16
Mean gene length (bp)	1,189
Per cent coding	44.1
Per cent genes with introns	6.2
tRNA number	33

The Genome of the Basidiomycetous Yeast and Human Pathogen *Cryptococcus neoformans*

Cryptococcus neoformans is a basidiomycetous yeast ubiquitous in the environment, a model for fungal pathogenesis, and an opportunistic human pathogen of global importance. We have sequenced its ~20-megabase genome, which contains ~6500 intron-rich gene structures and encodes a transcriptome abundant in alternatively spliced and antisense messages. The genome is rich in transposons, many of which cluster at candidate centromeric regions. The presence of these transposons may drive karyotype instability and phenotypic variation. *C. neoformans* encodes unique genes that may contribute to its unusual virulence properties, and comparison of two phenotypically distinct strains reveals variation in gene content in addition to sequence polymorphisms between the genomes.

EUKARYOTIC CELL, Jan. 2011, p. 34–42 Vol. 10, No. 1
1535-9778/11/$12.00 doi:10.1128/EC.00242-10

Comparative Genomics and the Evolution of Pathogenicity in Human Pathogenic Fungi[▽]

Gary P. Moran, David C. Coleman, and Derek J. Sullivan*

Microbiology Research Unit, Division of Oral Biosciences, Dublin Dental University Hospital, Trinity College Dublin, University of Dublin, Dublin 2, Republic of Ireland

Because most fungi have evolved to be free-living in the environment and because the infections they cause are usually opportunistic in nature, it is often difficult to identify specific traits that contribute to fungal pathogenesis. In recent years, there has been a surge in the number of sequenced genomes of human fungal pathogens, and comparison of these sequences has proved to be an excellent resource for exploring commonalities and differences in how these species interact with their hosts. In order to survive in the human body, fungi must be able to adapt to new nutrient sources and environmental stresses. Therefore, genes involved in carbohydrate and amino acid metabolism and transport and genes encoding secondary metabolites tend to be overrepresented in pathogenic species (e.g., *Aspergillus fumigatus*). However, it is clear that human commensal yeast species such as *Candida albicans* have also evolved a range of specific factors that facilitate direct interaction with host tissues. The evolution of virulence across the human pathogenic fungi has occurred largely through very similar mechanisms. One of the most important mechanisms is gene duplication and the expansion of gene families, particularly in subtelomeric regions. Unlike the case for prokaryotic pathogens, horizontal transfer of genes between species and other genera does not seem to have played a significant role in the evolution of fungal virulence. New sequencing technologies promise the prospect of even greater numbers of genome sequences, facilitating the sequencing of multiple genomes and transcriptomes within individual species, and will undoubtedly contribute to a deeper insight into fungal pathogenesis.

Draft Genome of the Filarial Nematode Parasite *Brugia malayi*

Parasitic nematodes that cause elephantiasis and river blindness threaten hundreds of millions of people in the developing world. We have sequenced the ~90 megabase (Mb) genome of the human filarial parasite *Brugia malayi* and predict ~11,500 protein coding genes in 71 Mb of robustly assembled sequence. Comparative analysis with the free-living, model nematode *Caenorhabditis elegans* revealed that, despite these genes having maintained little conservation of local synteny during ~350 million years of evolution, they largely remain in linkage on chromosomal units. More than 100 conserved operons were identified. Analysis of the predicted proteome provides evidence for adaptations of *B. malayi* to niches in its human and vector hosts and insights into the molecular basis of a mutualistic relationship with its *Wolbachia* endosymbiont. These findings offer a foundation for rational drug design.

Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism

Charles H. Opperman[a,b,c], David M. Bird[a,b], Valerie M. Williamson[d], Dan S. Rokhsar[e], Mark Burke[a], Jonathan Cohn[a], John Cromer[a], Steve Diener[a,f], Jim Gajan[a], Steve Graham[a], T. D. Houfek[a,g], Qingli Liu[d,h], Therese Mitros[i], Jennifer Schaff[a,j], Reenah Schaffer[a], Elizabeth Scholl[a], Bryon R. Sosinski[k,l], Varghese P. Thomas[d], and Eric Windham[a]

We have established *Meloidogyne hapla* as a tractable model plant-parasitic nematode amenable to forward and reverse genetics, and we present a complete genome sequence. At 54 Mbp, *M. hapla* represents not only the smallest nematode genome yet completed, but also the smallest metazoan, and defines a platform to elucidate mechanisms of parasitism by what is the largest uncontrolled group of plant pathogens worldwide. The *M. hapla* genome encodes significantly fewer genes than does the free-living nematode *Caenorhabditis elegans* (most notably through a reduction of odorant receptors and other gene families), yet it has acquired horizontally from other kingdoms numerous genes suspected to be involved in adaptations to parasitism. In some cases, amplification and tandem duplication have occurred with genes suspected of being acquired horizontally and involved in parasitism of plants. Although *M. hapla* and *C. elegans* diverged >500 million years ago, many developmental and biochemical pathways, including those for dauer formation and RNAi, are conserved. Although overall genome organization is not conserved, there are areas of microsynteny that may suggest a primary biological function in nematodes for those genes in these areas. This sequence and map represent a wealth of biological information on both the nature of nematode parasitism of plants and its evolution.

The genome of the heartworm, *Dirofilaria immitis*, reveals drug and vaccine targets

ABSTRACT The heartworm *Dirofilaria immitis* is an important parasite of dogs. Transmitted by mosquitoes in warmer climatic zones, it is spreading across southern Europe and the Americas at an alarming pace. There is no vaccine, and chemotherapy is prone to complications. To learn more about this parasite, we have sequenced the genomes of *D. immitis* and its endosymbiont *Wolbachia*. We predict 10,179 protein coding genes in the 84.2 Mb of the nuclear genome, and 823 genes in the 0.9-Mb *Wolbachia* genome. The *D. immitis* genome harbors neither DNA transposons nor active retrotransposons, and there is very little genetic variation between two sequenced isolates from Europe and the United States. The differential presence of anabolic pathways such as heme and nucleotide biosynthesis hints at the intricate metabolic interrelationship between the heartworm and *Wolbachia*. Comparing the proteome of *D. immitis* with other nematodes and with mammalian hosts, we identify families of potential drug targets, immune modulators, and vaccine candidates. This genome sequence will support the development of new tools against dirofilariasis and aid efforts to combat related human pathogens, the causative agents of lymphatic filariasis and river blindness.—Godel, C.,

TABLE 2. *Candidate drug targets, top-down search: current anthelmintics and their known targets in* C. elegans *and orthologs in* D. immitis

Chemical class	Drug	Target	*C. elegans*	*D. immitis*
Benzimidazole	Albendazole Flubendazole Mebendazole	β-Tubulin	BEN-1	DIMM36740
Imidazothiazole	Levamisole	nACh receptor	LEV-1	
			LEV-8	
			UNC-29	DIMM30000
			UNC-38	DIMM45965
			UNC-63	DIMM08405
Macrocyclic lactone	Ivermectin	Glutamate receptor	AVR-14	DIMM16610
	Milbemycin		AVR-15	
	Moxidectin		GLC-1	
	Selamectin		GLC-2	DIMM25280, DIMM21120
			GLC-3	
			GLC-4	DIMM22030
		GABA receptor	EXP-1	DIMM57890
			GAB-1	
			UNC-49	DIMM33210
Cyclodepsipeptide	Emodepside	K^+ channel	SLO-1	DIMM33710
		Latrophilin GPCR	LAT-1	DIMM37270, DIMM37275
			LAT-2	DIMM17690
Aminoacetonitrile derivative	Monepantel	nACh receptor	ACR-23	
			DES-2	

nAChR, nicotinic acetylcholine; GPCR, G protein-coupled receptor.

Genome sequence of the human malaria parasite *Plasmodium falciparum*

The parasite *Plasmodium falciparum* is responsible for hundreds of millions of cases of malaria, and kills more than one million African children annually. Here we report an analysis of the genome sequence of *P. falciparum* clone 3D7. The 23-megabase nuclear genome consists of 14 chromosomes, encodes about 5,300 genes, and is the most (A + T)-rich genome sequenced to date. Genes involved in antigenic variation are concentrated in the subtelomeric regions of the chromosomes. Compared to the genomes of free-living eukaryotic microbes, the genome of this intracellular parasite encodes fewer enzymes and transporters, but a large proportion of genes are devoted to immune evasion and host–parasite interactions. Many nuclear-encoded proteins are targeted to the apicoplast, an organelle involved in fatty-acid and isoprenoid metabolism. The genome sequence provides the foundation for future studies of this organism, and is being exploited in the search for new drugs and vaccines to fight malaria.

Table 1 ***Plasmodium falciparum*** **nuclear genome summary and comparison to other organisms**

Feature	Value				
	P. falciparum	*S. pombe*	*S. cerevisiae*	*D. discoideum*	*A. thaliana*
Size (bp)	22,853,764	12,462,637	12,495,682	8,100,000	115,409,949
(G + C) content (%)	19.4	36.0	38.3	22.2	34.9
No. of genes	5,268*	4,929	5,770	2,799	25,498
Mean gene length† (bp)	2,283	1,426	1,424	1,626	1,310
Gene density (bp per gene)	4,338	2,528	2,088	2,600	4,526
Per cent coding	52.6	57.5	70.5	56.3	28.8
Genes with introns (%)	53.9	43	5.0	68	79
Exons					
Number	12,674	ND	ND	6,398	132,982
No. per gene	2.39	ND	NA	2.29	5.18
(G + C) content (%)	23.7	39.6	28.0	28.0	ND
Mean length (bp)	949	ND	ND	711	170
Total length (bp)	12,028,350	ND	ND	4,548,978	33,249,250
Introns					
Number	7,406	4,730	272	3,587	107,784
(G + C) content (%)	13.5	ND	NA	13.0	ND
Mean length (bp)	178.7	81	NA	177	170
Total length (bp)	1,323,509	383,130	ND	643,899	18,055,421
Intergenic regions					
(G + C) content (%)	13.6	ND	ND	14.0	ND
Mean length (bp)	1,694	952	515	786	ND
RNAs					
No. of tRNA genes	43	174	ND	73	ND
No. of 5S rRNA genes	3	30	ND	NA	ND
No. of 5.8S, 18S and 28S rRNA units	7	200–400	ND	NA	700–800

ND, not determined; NA, not applicable. 'No. of genes' for *D. discoideum* are for chromosome 2 (ref. 155) and in some cases represent extrapolations to the entire genome. Sources of data for the other organisms: *S. pombe*[65], *S. cerevisiae*[156], *D. discoideum*[155] and *A. thaliana*[157].
*70% of these genes matched expressed sequence tags or encoded proteins detected by proteomics analyses[14,15].
†Excluding introns.

The *P. falciparum*, *Anopheles gambiae* and *Homo sapiens* genome sequences have been completed in the past two years, and represent new starting points in the centuries-long search for solutions to the malaria problem. For the first time, a wealth of information is available for all three organisms that comprise the life cycle of the malaria parasite, providing abundant opportunities for the study of each species and their complex interactions that result in disease. The rapid pace of improvements in sequencing technology and the declining costs of sequencing have made it possible to begin genome sequencing efforts for *Plasmodium vivax*, the second major human malaria parasite, several malaria parasites of animals, and for many related parasites such as *Theileria* and *Toxoplasma*. These will be extremely useful for comparative purposes. Last, this technology will enable sampling of parasite, vector and host genomes in the field, providing information to support the development, deployment and monitoring of malaria control methods.

In the short term, however, the genome sequences alone provide little relief to those suffering from malaria. The work reported here and elsewhere needs to be accompanied by larger efforts to develop new methods of control, including new drugs and vaccines, improved diagnostics and effective vector control techniques. Much remains to be done. Clearly, research and investments to develop and implement new control measures are needed desperately if the social and economic impacts of malaria are to be relieved. The increased attention given to malaria (and to other infectious diseases affecting tropical countries) at the highest levels of government, and the initiation of programmes such as the Global Fund to Fight AIDS, Tuberculosis and Malaria[148], the Multilateral Initiative on Malaria in Africa[149], the Medicines for Malaria Venture[150], and the Roll Back Malaria campaign[151], provide some hope of progress in this area. It is our hope and expectation that researchers around the globe will use the information and biological insights provided by complete genome sequences to accelerate the search for solutions to diseases affecting the most vulnerable of the world's population. □

Comparative Genomics of Trypanosomatid Parasitic Protozoa

A comparison of gene content and genome architecture of *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania major*, three related pathogens with different life cycles and disease pathology, revealed a conserved core proteome of about 6200 genes in large syntenic polycistronic gene clusters. Many species-specific genes, especially large surface antigen families, occur at nonsyntenic chromosome-internal and subtelomeric regions. Retroelements, structural RNAs, and gene family expansion are often associated with syntenic discontinuities that—along with gene divergence, acquisition and loss, and rearrangement within the syntenic regions—have shaped the genomes of each parasite. Contrary to recent reports, our analyses reveal no evidence that these species are descended from an ancestor that contained a photosynthetic endosymbiont.

RESEARCH ARTICLE

The Genome Sequence of *Trypanosoma cruzi*, Etiologic Agent of Chagas Disease

Whole-genome sequencing of the protozoan pathogen *Trypanosoma cruzi* revealed that the diploid genome contains a predicted 22,570 proteins encoded by genes, of which 12,570 represent allelic pairs. Over 50% of the genome consists of repeated sequences, such as retrotransposons and genes for large families of surface molecules, which include trans-sialidases, mucins, gp63s, and a large novel family (>1300 copies) of mucin-associated surface protein (MASP) genes. Analyses of the *T. cruzi*, *T. brucei*, and *Leishmania major* (Tritryp) genomes imply differences from other eukaryotes in DNA repair and initiation of replication and reflect their unusual mitochondrial DNA. Although the Tritryp lack several classes of signaling molecules, their kinomes contain a large and diverse set of protein kinases and phosphatases; their size and diversity imply previously unknown interactions and regulatory processes, which may be targets for intervention.

Table 1. General features of the Tritryp genomes. We found 5812 syntenic three-way COGs and 346 nonsyntenic three-way COGs. Mbp, mega–base pairs; NC, not computed.

	T. brucei	*T. cruzi*	*L. major*
Haploid genome size (Mbp)	25*	55	33
No. of chromosomes (per haploid genome)	11*	~28†	36
No. of genes (per haploid genome)	9068‡	~12,000§	8311‖
Total regions with synteny blocks (Mbp)	19.9	NC	30.7
Mean CDS size (bp) in syntenic three-way COGs	1511	1457	1731
Mean inter-CDS size (bp) between syntenic three-way COGs	721	561	1431

*Excluding ~100 mini- and intermediate-sized chromosomes (totaling ~10 Mb). and homologs can differ substantially in size. ‡Includes 904 pseudogenes. has not been determined in *T. cruzi*. ‖Includes 34 pseudogenes. †The exact number is not known §The exact number of haploid genes

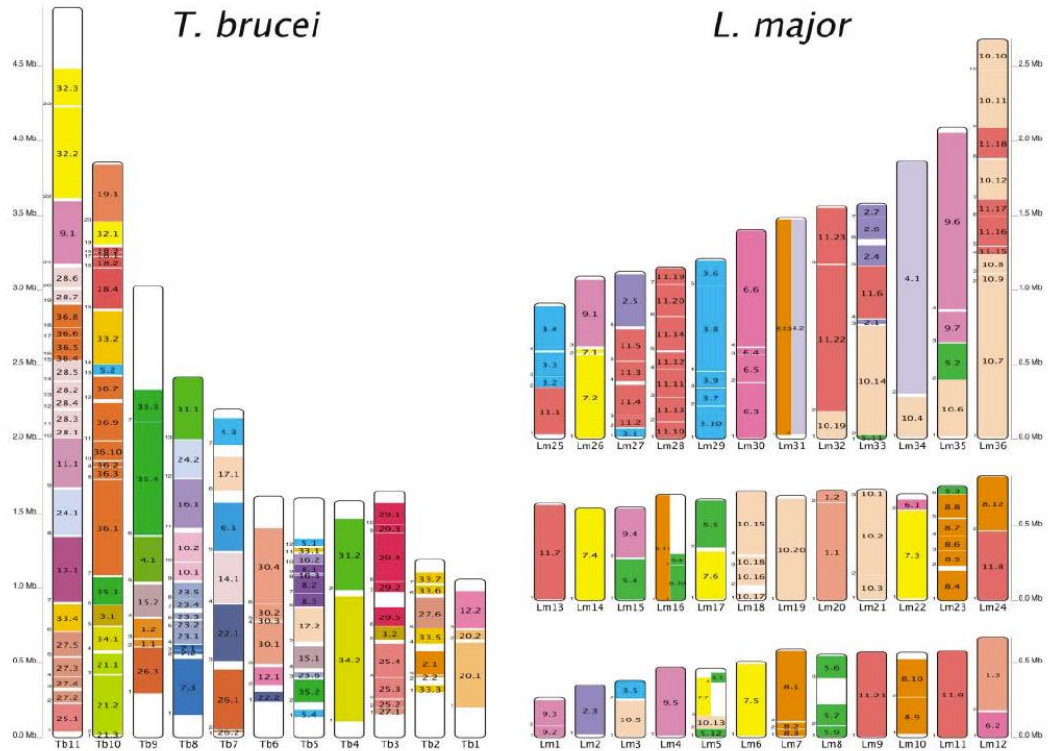
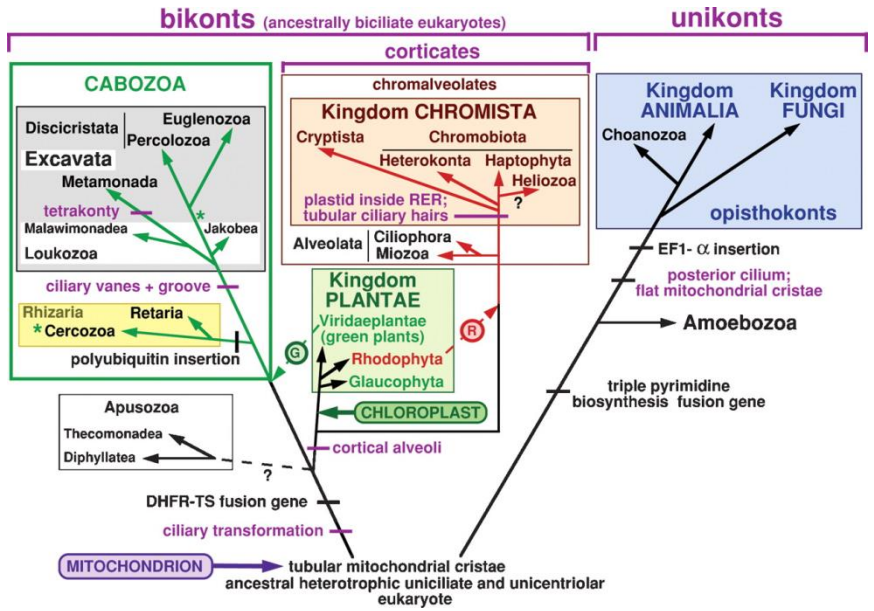


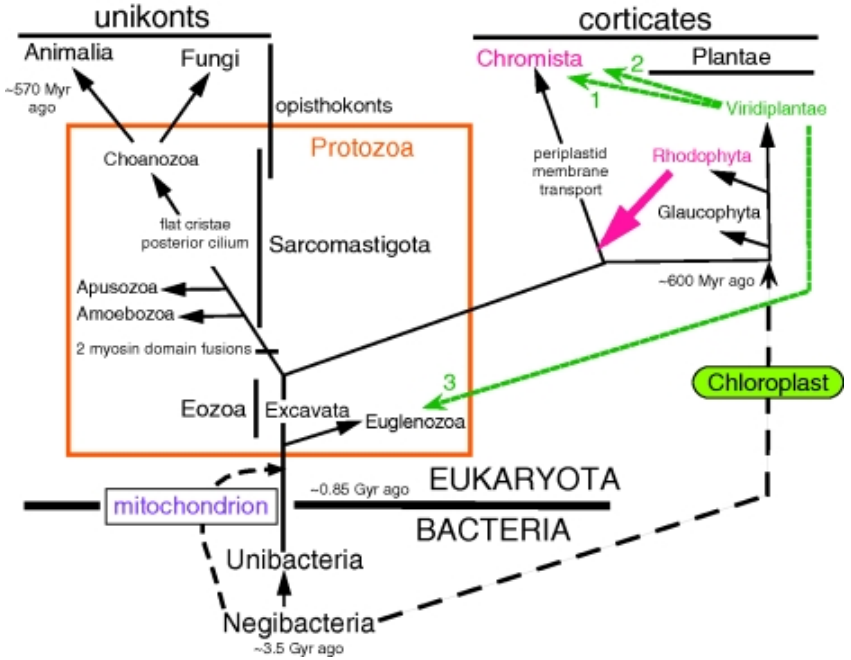
Fig. 2. Synteny maps. The 36 different colors in the *T. brucei* (left) panel represent the locations of the indicated synteny blocks in the 36 chromosomes of *L. major*, and the 11 colors in the *L. major* (right) panel depict the locations of the indicated synteny blocks in the 11 chromosomes of *T. brucei*. Each synteny block is named using a double nomenclature that refers to the chromosomal location of the block in both species. Labels on the left outside margin of the synteny blocks denote the block number in the reference genome. Labels within synteny blocks refer to their location on the other genome. For example, synteny block Tb1.1 of *T. brucei* chromosome 1 (Tb1; lower right of left panel) is synteny block Lm20.1 of *L. major* chromosome 20 (Lm20). As another example, all of the yellow synteny blocks in the *L. major* panel (blocks Tb7.1 to Tb7.7) are on *T. brucei* chromosome 7 (Tb7). Synteny blocks are defined as groups of five or more *T. brucei* genes that possess an ortholog on the same *L. major* chromosome (*2*). The entire map contains 7974 *T. brucei* and 7466 *L. major* protein-coding genes in 110 synteny blocks. Plate 1 and fig. S3, A to K, show more detailed views and table S4, A and B, has complete lists of genes and block coordinates.

Kaj je vseboval genom najstarejših/prvotnih eukariontov



Narobe !!

Stechmann A, Cavalier-Smith T. **Rooting the eukaryote tree by using a derived gene fusion.** Science 2002 Jul 5;297(5578):89-91.



Nov pogled, bolj pravilen

Cavalier-Smith T. **Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree**. Biol Lett. 2010 Jun 23;6(3):342-345.

The Genome of *Naegleria gruberi* Illuminates Early Eukaryotic Versatility

Genome sequences of diverse free-living protists are essential for understanding eukaryotic evolution and molecular and cell biology. The free-living amoeboflagellate *Naegleria gruberi* belongs to a varied and ubiquitous protist clade (Heterolobosea) that diverged from other eukaryotic lineages over a billion years ago. Analysis of the 15,727 protein-coding genes encoded by *Naegleria*'s 41 Mb nuclear genome indicates a capacity for both aerobic respiration and anaerobic metabolism with concomitant hydrogen production, with fundamental implications for the evolution of organelle metabolism. The *Naegleria* genome facilitates substantially broader phylogenomic comparisons of free-living eukaryotes than previously possible, allowing us to identify thousands of genes likely present in the pan-eukaryotic ancestor, with 40% likely eukaryotic inventions. Moreover, we construct a comprehensive catalog of amoeboid-motility genes. The *Naegleria* genome, analyzed in the context of other protists, reveals a remarkably complex ancestral eukaryote with a rich repertoire of cytoskeletal, sexual, signaling, and metabolic modules.

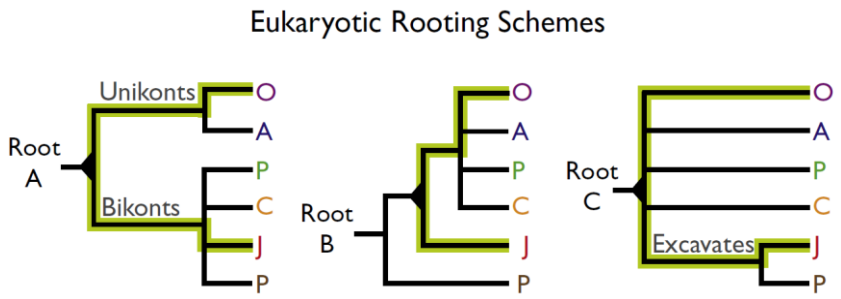
Table 1. Genome Statistics from *Naegleria gruberi* and Selected Species

Species	Genome Size (Mbp)	No. Chromosomes	%GC	Protein-Coding Loci	% Coding	% Genes w/ Introns	Introns per Gene	Median Intron Length (bp)
Naegleria	41	> = 12	33	15, 727	57.8	36	0.7	60
Human	2851	23	41	23, 328	1.2	83	7.8	20, 383
Neurospora	40	7	54	10, 107	36.4	80	1.7	72
Dictyostelium	34	6	22	13, 574	62.2	68	1.3	236
Arabidopsis	140.1	5	36	26, 541	23.7	80	4.4	55
Chlamydomonas	121	17	64	14, 516	16.3	91	7.4	174
T. brucei	26.1	>100	46	9152	52.6	~0 (1 total)	ND	ND
Giardia	11.7	5	49	6480	71.4	~0 (4 total)	ND	ND

See also Figure S1 and Tables S1, S2, S3, S8, S9, and S10. ND, not determined.

Inferring the Protein Complement of the Eukaryotic Ancestor

What genes were present in the common ancestor of all eukaryotes? Prior inventories of ancestral eukaryotic genes have been based on two or three eukaryotic groups (Hartman and Fedorov, 2002; Tatusov et al., 2003). This limited sampling, and the limited availability of free-living protist genome sequences, may have significantly underestimated the protein complement of the eukaryotic common ancestor. We used 17 genomes from all six major groups and constructed 4133 ancient eukaryotic gene families, requiring (1) a minimum of one *Naegleria* protein and two orthologs and (2) one ortholog from another major eukaryotic group. These ancient gene families are conceptually similar to KOGs (eukaryotic clusters of orthologous groups), which were

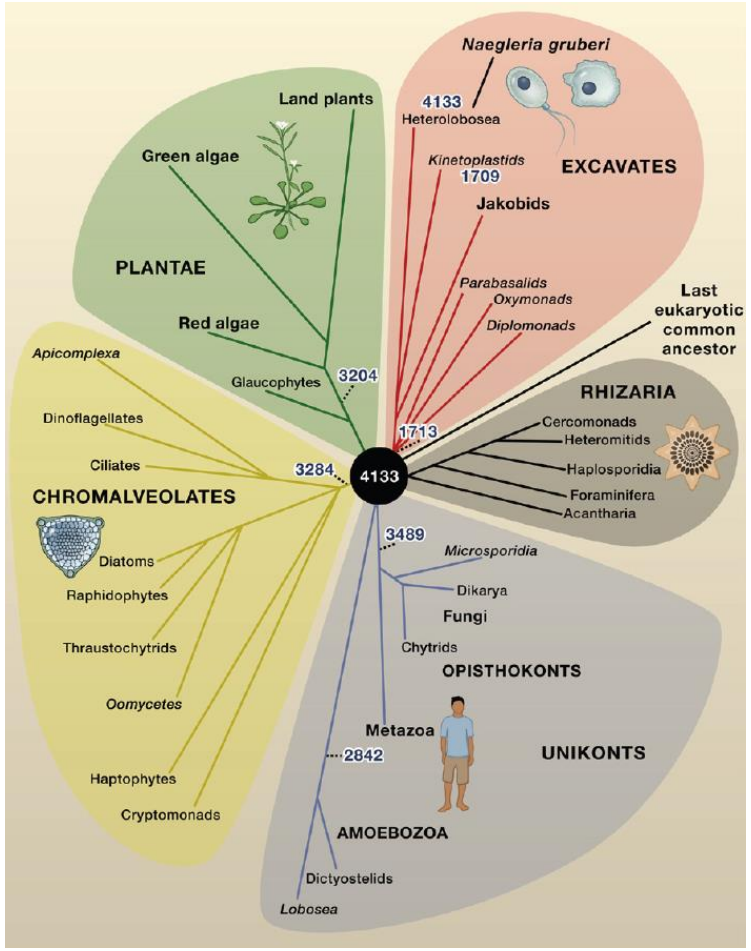


depicts three contending hypotheses for the root. Root A: early divergence of unikonts and bikonts (Stechmann and Cavalier-Smith, 2002). Root B: the largely parasitic POD lineage branching first, followed by JEH (including *Naegleria*) (Ciccarelli et al., 2006). Root C: POD and JEH uniting to form the "Excavates" (Supplemental Information). The branches connecting *Naegleria* to humans are highlighted in green, with a black triangle indicating their last common ancestor. See also Text S1.

The Incredible Expanding Ancestor of Eukaryotes

Eugene V. Koonin[1,*]
[1]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA
*Correspondence: koonin@ncbi.nlm.nih.gov
DOI 10.1016/j.cell.2010.02.022

Comparing the genome sequences of free-living organisms in the five eukaryotic supergroups enables predictions to be made about the genome of the last common ancestor of eukaryotes. The genome sequence of the amoeboflagellate *Naegleria gruberi* reported by Fritz-Laylin et al. (2010) reveals the surprising complexity of this unicellular organism and, by inference, of the last common eukaryotic ancestor.



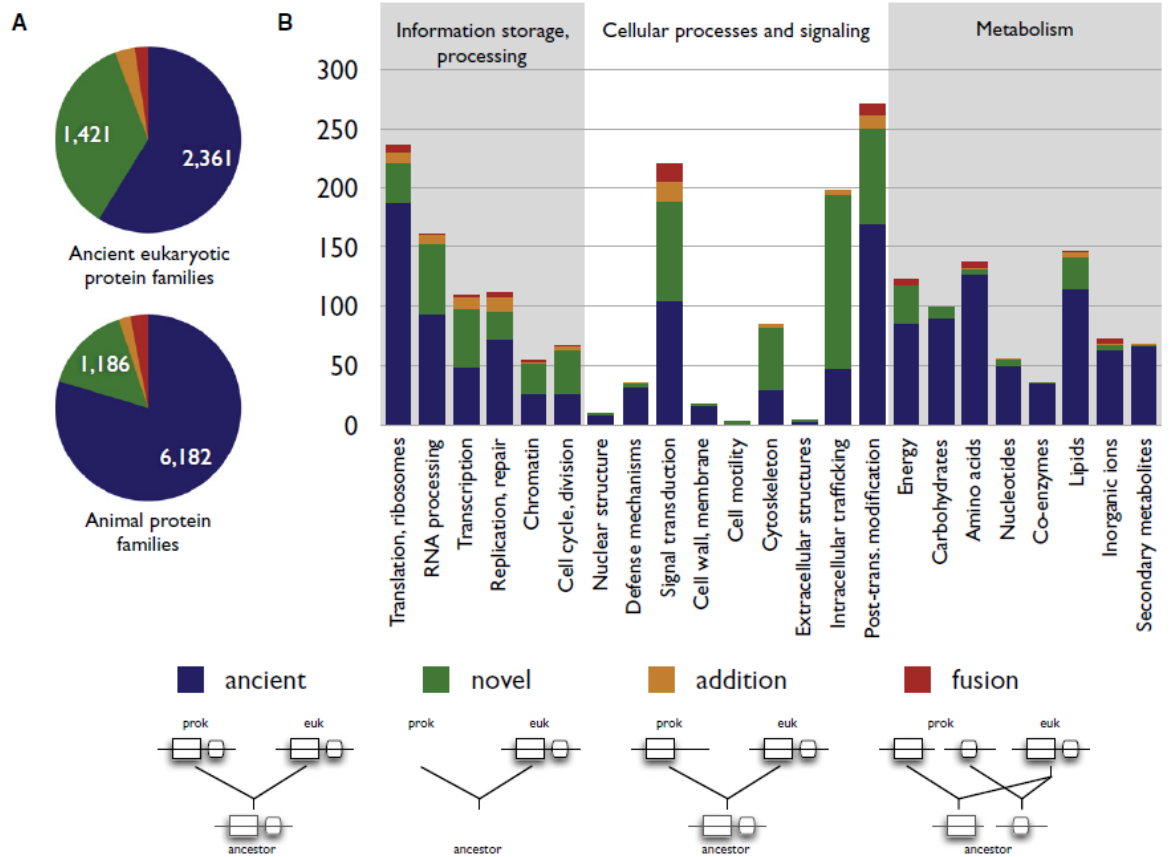


Figure 6. Ancient Origin and Innovation in Eukaryotic Proteins
Schematics of the four scenarios of protein origin we consider are along the bottom and color-coded in the charts: ancient (blue), novel (green), addition of a eukaryote-specific protein domain (orange), and eukaryotic-specific fusion of two domains (red). The protein families that could be categorized are presented in (A) overview pie charts comparing the origins of protein families in ancient eukaryotes (top) and animals (bottom, from Putnam et al., 2007) and (B) stacked barcharts showing subsets of the ancient eukaryotic families divided by KOG function, omitting unknown and general KOG functions. prok, prokaryotic (i.e., archaeal and/or bacterial); euk, eukaryotic; Post-trans., post-translational. See also Tables S4, S19, S20, S21, S22, and S23.

Figure 1. Eukaryotic Evolution
The five eukaryotic supergroups—Excavates, Rhizaria, Unikonts, Chromalveolates, and Plantae—are shown to diverge directly from the last common ancestor (black circle) because the relationship between individual supergroups is uncertain (Keeling, 2007). Analysis of the genome sequence of the free-living amoeboflagellate *Naegleria gruberi* in the Excavate supergroup reveals that this organism has 4133 genes that are shared by at least one other eukaryote supergroup (Fritz-Laylin et al., 2010). These 4133 genes are inferred to be ancestral genes that were present in the last common ancestor of eukaryotes, suggesting that this common ancestor was surprisingly complex. The numbers of putative ancestral genes present in selected major clades are indicated in blue. The Rhizaria supergroup is included for completeness, despite the current absence of sequenced genomes. The names of groups that include mostly parasites are italicized. Branch lengths are arbitrary.

Genomi arhaičnih ljudi (Neandertalci in Denizovanci)

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1]*†‡ Johannes Krause,[1]†§ Adrian W. Briggs,[1]†§ Tomislav Maricic,[1]†§ Udo Stenzel,[1]†§ Martin Kircher,[1]†§ Nick Patterson,[2]†§ Heng Li,[2]† Weiwei Zhai,[3]†|| Markus Hsi-Yang Fritz,[4]† Nancy F. Hansen,[5]† Eric Y. Durand,[3]† Anna-Sapfo Malaspinas,[3]† Jeffrey D. Jensen,[6]† Tomas Marques-Bonet,[7,13]† Can Alkan,[7]† Kay Prüfer,[1]† Matthias Meyer,[1]† Hernán A. Burbano,[1]† Jeffrey M. Good,[1,8]† Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14]¶ Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18]† Evan E. Eichler,[7]† Daniel Falush,[19]† Ewan Birney,[4]† James C. Mullikin,[5]† Montgomery Slatkin,[3]† Rasmus Nielsen,[3]† Janet Kelso,[1]† Michael Lachmann,[1]† David Reich,[2,20]*† Svante Pääbo[1]*†

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

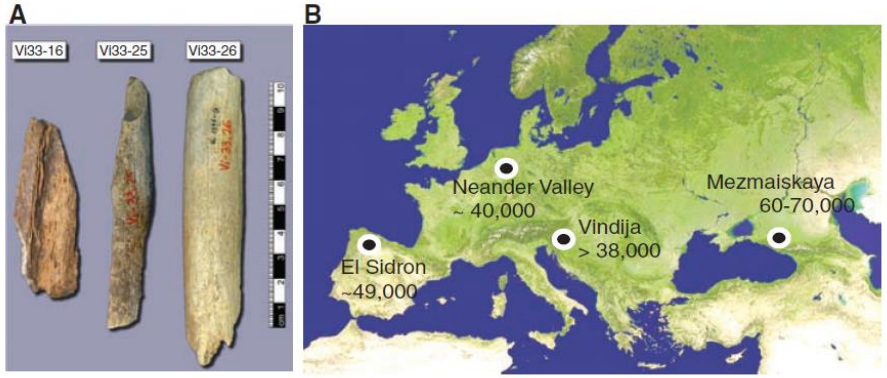


Fig. 1. Samples and sites from which DNA was retrieved. (**A**) The three bones from Vindija from which Neandertal DNA was sequenced. (**B**) Map showing the four archaeological sites from which bones were used and their approximate dates (years B.P.).

A High-Coverage Genome Sequence from an Archaic Denisovan Individual

Matthias Meyer,[1]*‡ Martin Kircher,[1]*† Marie-Theres Gansauge,[1] Heng Li,[2] Fernando Racimo,[1] Swapan Mallick,[2,3] Joshua G. Schraiber,[4] Flora Jay,[4] Kay Prüfer,[1] Cesare de Filippo,[1] Peter H. Sudmant,[6] Can Alkan,[5,6] Qiaomei Fu,[1,7] Ron Do,[2] Nadin Rohland,[2,3] Arti Tandon,[2,3] Michael Siebauer,[1] Richard E. Green,[8] Katarzyna Bryc,[3] Adrian W. Briggs,[3] Udo Stenzel,[1] Jesse Dabney,[1] Jay Shendure,[6] Jacob Kitzman,[6] Michael F. Hammer,[9] Michael V. Shunkov,[10] Anatoli P. Derevianko,[10] Nick Patterson,[2] Aida M. Andrés,[1] Evan E. Eichler,[6,11] Montgomery Slatkin,[4] David Reich,[2,3]‡ Janet Kelso,[1] Svante Pääbo[1]‡

We present a DNA library preparation method that has allowed us to reconstruct a high-coverage (30×) genome sequence of a Denisovan, an extinct relative of Neandertals. The quality of this genome allows a direct estimation of Denisovan heterozygosity indicating that genetic diversity in these archaic hominins was extremely low. It also allows tentative dating of the specimen on the basis of "missing evolution" in its genome, detailed measurements of Denisovan and Neandertal admixture into present-day human populations, and the generation of a near-complete catalog of genetic changes that swept to high frequency in modern humans since their divergence from Denisovans.

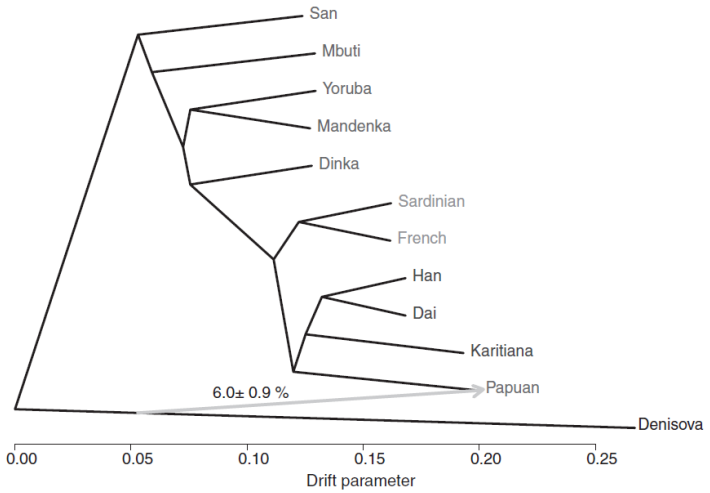
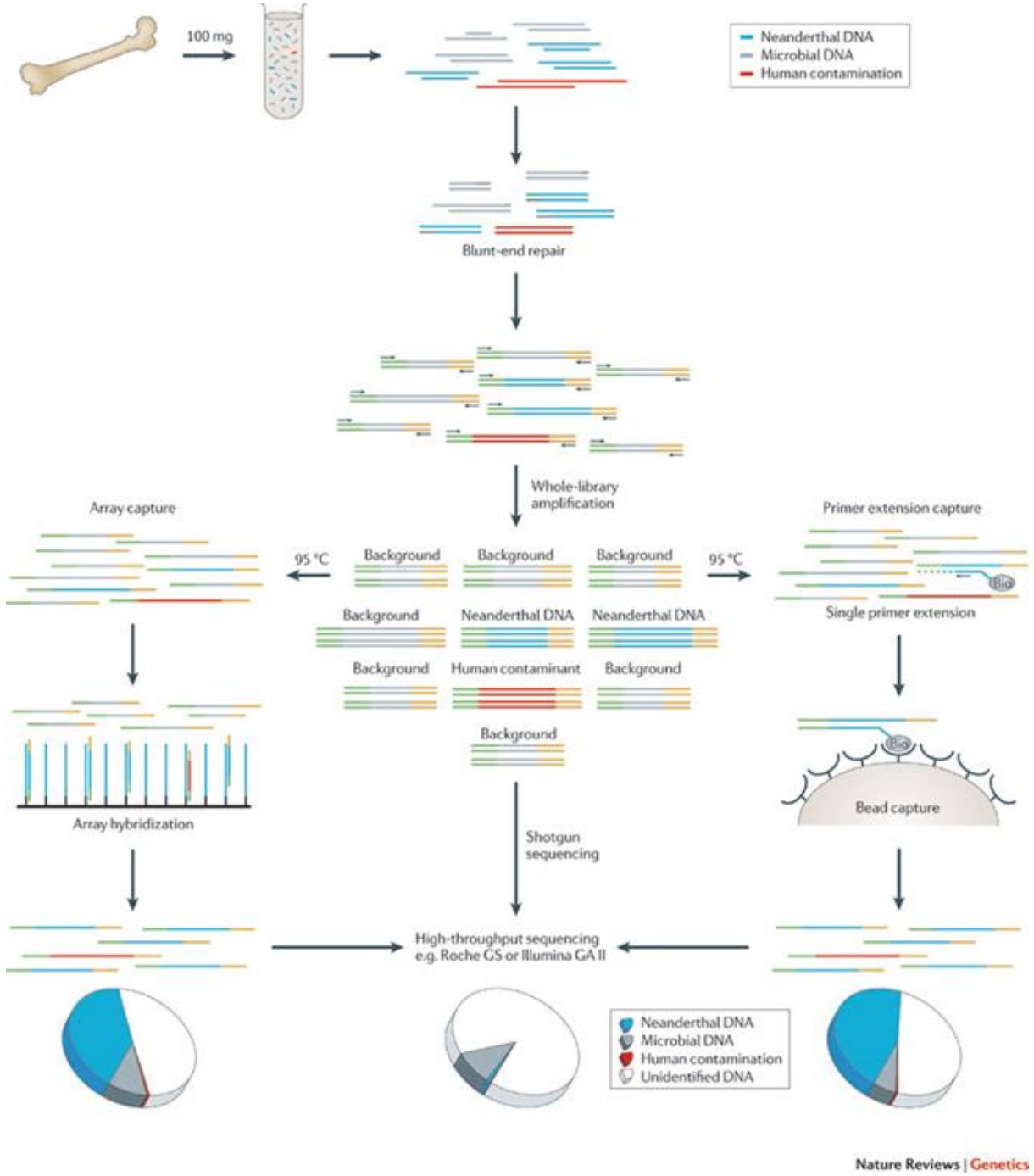


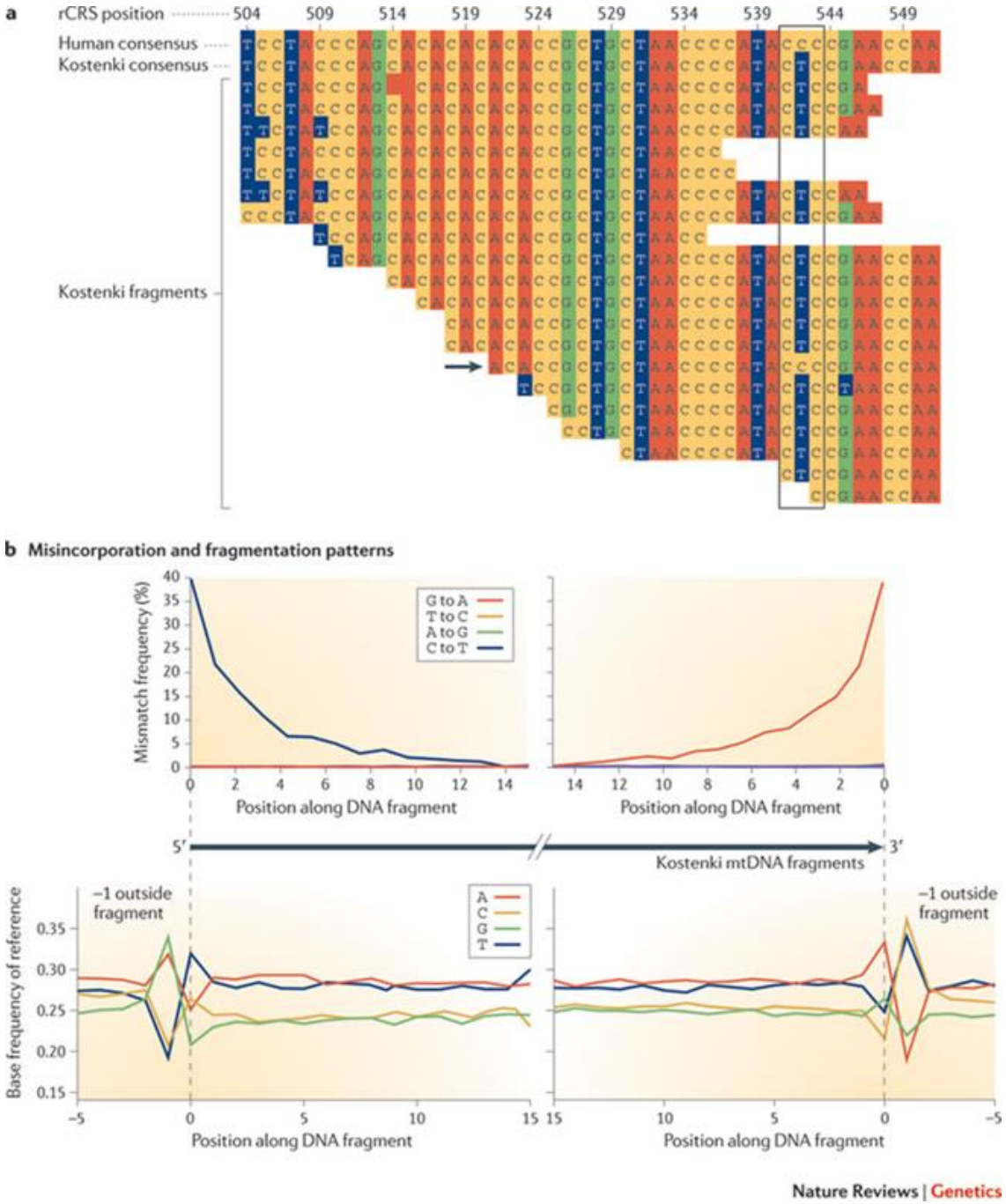
Fig. 3. Maximum likelihood tree relating the Denisovan genome and the genomes of 11 present-day humans, allowing one migration event (shown as a gray arrow).

Implications for archaic and modern human history. It is striking that genetic diversity among Denisovans was low although they were present in Siberia as well as presumably in Southeast Asia where they interacted with the ancestors of present-day Melanesians (*6*). Only future research can show how wide their geographic range was at any one time in their history. However, it is likely that they have expanded from a small population size with not enough time elapsing for genetic diversity to correspondingly increase. When technical improvements such as the one presented here will make it possible to sequence a Neandertal genome to a quality comparable to the Denisovan and modern genomes, it will be important to clarify whether the temporal trajectory of Neandertal effective population size matches that of the Denisovans. If that is the case, it is likely that the low Denisovan diversity reflects the expansion out of Africa of a population ancestral to both Denisovans and Neandertals, a possibility that seems compatible with the dates for population divergences and population size changes presented.



High-throughput sequencing of ancient DNA

A schematic representation of high-throughput sequencing of DNA from fossil remains, here depicted as a Neanderthal bone. The ancient DNA is first blunt-end repaired, and then DNA adaptors are added to each end. The final product, called the sequencing library, serves as the input for various high-throughput sequencing strategies and technologies. All ancient DNA molecules in the library will be first amplified using the adaptors as priming sites in PCR. Aliquots that contain copies of all original ancient DNA molecules can be directly sequenced on a high-throughput sequencer (centre panel) or used in targeted enrichment via array (left panel) or primer extension capture (right panel) methods. The pie charts illustrate the percentage of Neanderthal DNA obtained by each of these approaches.



Distinguishing ancient from modern DNA

a | Estimating contamination with modern DNA. Shown is a section of an alignment of the complete mitochondrial DNA (mtDNA) genome (total 16,570 positions) of an early modern human from the Kostenki site, Russia31. The positions are based on the revised Cambridge reference sequence (rCRS). The first line of the alignment shows the consensus sequence obtained from 311 worldwide modern human mtDNAs. The second line shows the consensus sequence for 10,664 mtDNA fragments retrieved from the 30,000-year-old Kostenki early modern human bone. To get an estimate of contamination with modern human DNA, positions were identified where more than 99% of 311 modern human mtDNAs are different from the Kostenki consensus sequence. All fragments that overlap such a position (boxed) and are different from the Kostenki consensus base are likely to be modern human contamination. Only one fragment (indicated by an arrow) is inconsistent, suggesting a very low level of contemporary modern human contamination (1 out of 16 fragments that overlap this position and 1 out of 77 for the complete Kostenki mtDNA data set).

b | The spatial distribution of DNA degradation patterns that are typical for ancient DNA, shown here for the mtDNA fragments from the Kostenki early modern human. The upper panel shows DNA mismatches to a reference sequence for all ancient mtDNA fragments: more than 40% of Cs are seen as Ts at the 5′ end of the mtDNA fragments (left) and more than 40% of Gs are seen as As at the 3′ end (right). The lower panel shows the base frequency of the reference sequence: left, purines (A and G) are in high frequency one base pair upstream of the 5′ end of the start of the mtDNA sequence; right, pyrimidines (C and T) are in high frequency one base pair downstream of the 3′ end of the mtDNA sequence. The presence of such patterns can be used to test the authenticity of ancient modern human DNA.

Genetic roots of the first Americans

The whole-genome sequence of a human associated with the earliest widespread culture in North America confirms the Asian ancestry of the Clovis people and their relatedness to present-day Native Americans. SEE LETTER P.225

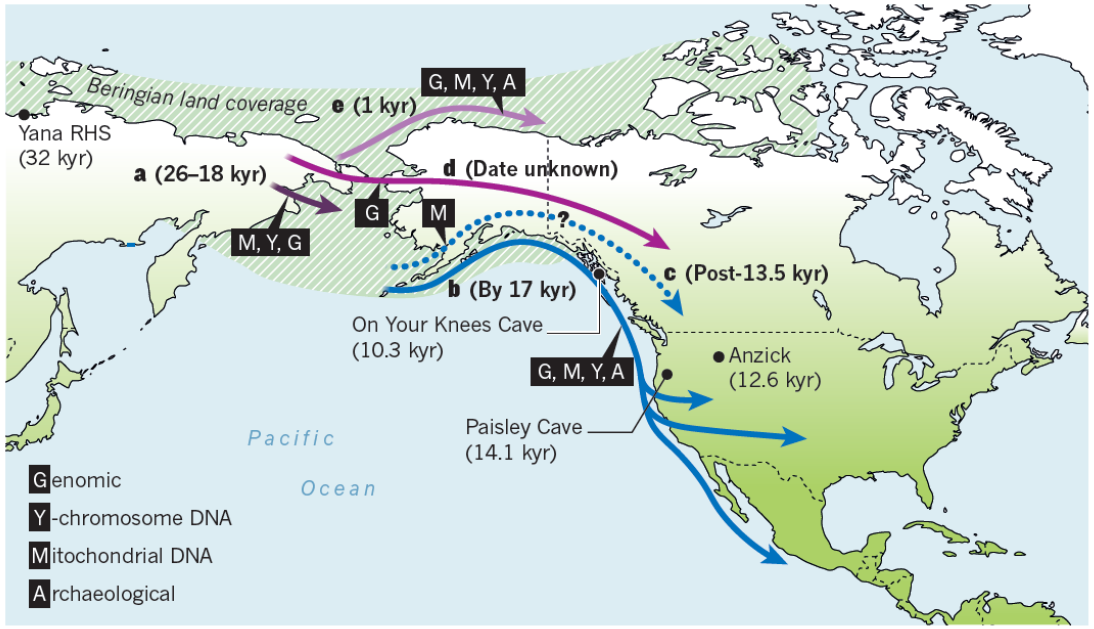


Figure 1 | Populating a continent. A hypothetical scenario for the peopling of the Americas, showing possible migration events (**a**–**e**) as described in the main text, coloured according to putative region of origin (Beringia, blue; Siberia, shades of purple). Also shown are some key archaeological sites in North America and Siberia, and the type of evidence (genomic, mitochondrial DNA, Y-chromosome DNA and archaeological) currently supporting each hypothesized migration. Shading depicts the extent of Beringia during the Last Glacial Maximum. (kyr, thousand years ago.)

Rasmussen M, et al. **The genome of a Late Pleistocene human from a Clovis burial site in western Montana.** Nature 2014 Feb 13;506(7487):225-229.

The genome of a Late Pleistocene human from a Clovis burial site in western Montana

Clovis, with its distinctive biface, blade and osseous technologies, is the oldest widespread archaeological complex defined in North America, dating from 11,100 to 10,700 14C years before present (BP) (13,000 to 12,600 calendar years BP). Nearly 50 years of archaeological research point to the Clovis complex as having developed south of the North American ice sheets from an ancestral technology. However, both the origins and the genetic legacy of the people who manufactured Clovis tools remain under debate. It is generally believed that these people ultimately derived from Asia and were directly related to contemporary Native Americans. An alternative, Solutrean, hypothesis posits that the Clovis predecessors emigrated from southwestern Europe during the Last Glacial Maximum. Here we report the genome sequence of a male infant (Anzick-1) recovered from the Anzick burial site in western Montana. The human bones date to 10,705 ± 35 14C years BP (approximately 12,707–12,556 calendar years BP) and were directly associated with Clovis tools. We sequenced the genome to an average depth of 14.4× and show that the **gene flow from the Siberian Upper Palaeolithic Mal'ta population into Native American ancestors is also shared by the Anzick-1 individual and thus happened before 12,600 years BP**. We also show that **the Anzick-1 individual is more closely related to all indigenous American populations than to any other group**. Our data are compatible with the **hypothesis that Anzick-1 belonged to a population directly ancestral to many contemporary Native Americans**. Finally, we find **evidence of a deep divergence in Native American populations that predates the Anzick-1 individual.**

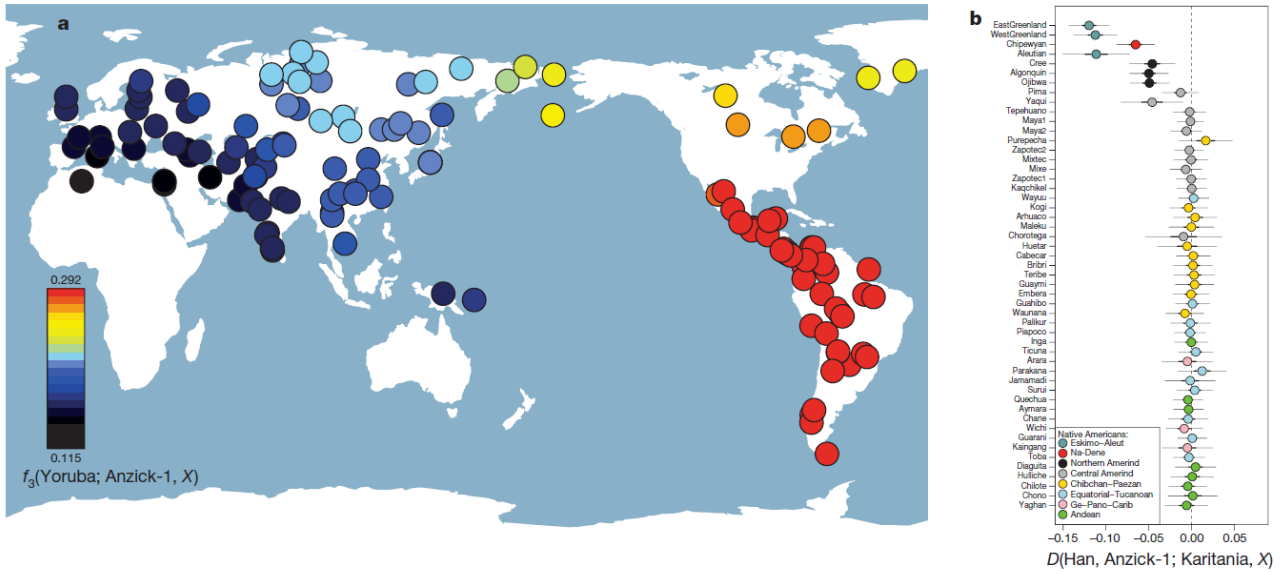


Figure 2 | Genetic affinity of Anzick-1. a, Anzick-1 is most closely related to Native Americans. Heat map representing estimated outgroup f_3-statistics for shared genetic history between the Anzick-1 individual and each of 143 contemporary human populations outside sub-Saharan Africa. **b**, Anzick-1 is less closely related to Northern Native American populations and a Yaqui individual than to Central and South Native Americans such as the Brazilian Karitiana. We computed *D*-statistics of the form *D*(Han, Anzick-1; Karitiana, *X*) to test the hypothesis that a second Native American population *X* is as closely related to Anzick-1 as the South American Karitiana is. Thick and thin whiskers represent 1 and 3 standard errors, respectively.

Genetska struktura človeških populacij (Amerika, Azija, Afrika, Židi itd.)

The Genetic Structure and History of Africans and African Americans

Sarah A. Tishkoff,[1,2]* Floyd A. Reed,[1]†‡ Françoise R. Friedlaender,[3]‡ Christopher Ehret,[4] Alessia Ranciaro,[1,2,5]§ Alain Froment,[6]§ Jibril B. Hirbo,[1,2] Agnes A. Awomoyi,[1]‖ Jean-Marie Bodo,[7] Ogobara Doumbo,[8] Muntaser Ibrahim,[9] Abdalla T. Juma,[9] Maritha J. Kotze,[10] Godfrey Lema,[11] Jason H. Moore,[12] Holly Mortensen,[1]¶ Thomas B. Nyambo,[11] Sabah A. Omar,[13] Kweli Powell,[1]# Gideon S. Pretorius,[14] Michael W. Smith,[15] Mahamadou A. Thera,[8] Charles Wambebe,[16] James L. Weber,[17] Scott M. Williams[18]

Africa is the source of all modern humans, but characterization of genetic variation and of relationships among populations across the continent has been enigmatic. We studied 121 African populations, four African American populations, and 60 non-African populations for patterns of variation at 1327 nuclear microsatellite and insertion/deletion markers. We identified 14 ancestral population clusters in Africa that correlate with self-described ethnicity and shared cultural and/or linguistic properties. We observed high levels of mixed ancestry in most populations, reflecting historical migration events across the continent. Our data also provide evidence for shared ancestry among geographically diverse hunter-gatherer populations (Khoesan speakers and Pygmies). The ancestry of African Americans is predominantly from Niger-Kordofanian (~71%), European (~13%), and other African (~8%) populations, although admixture levels varied considerably among individuals. This study helps tease apart the complex evolutionary history of Africans and African Americans, aiding both anthropological and genetic epidemiologic studies.

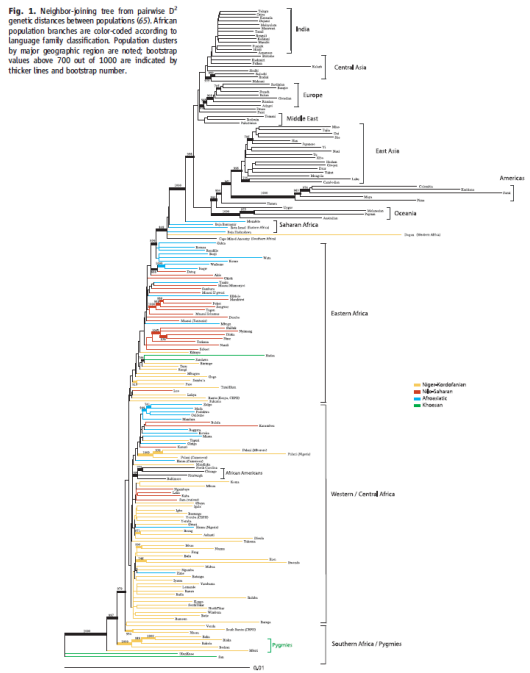
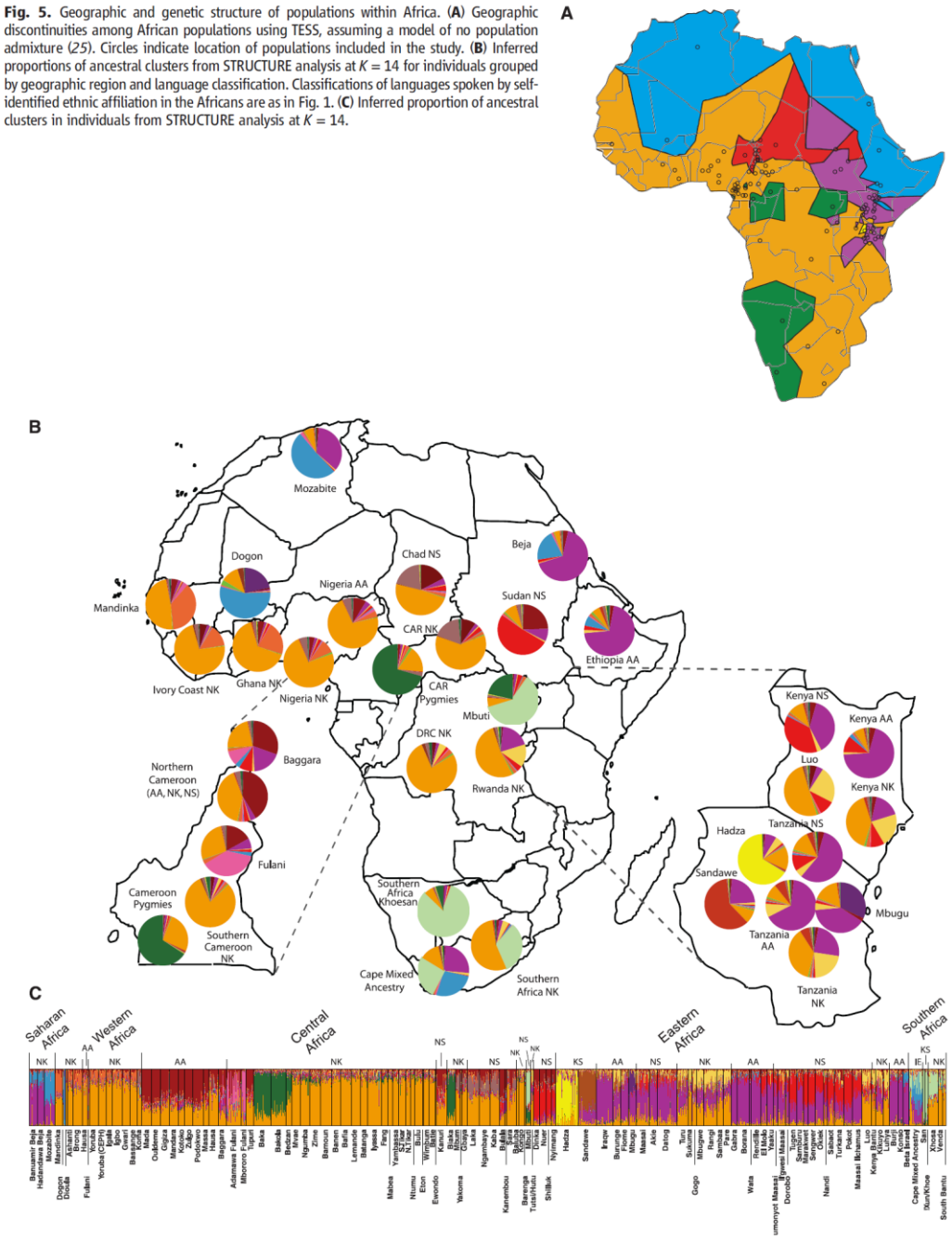


Fig. 1. Neighbor-joining tree from pairwise D^2 genetic distances between populations (65). African population branches are color-coded according to language family classification. Population clusters by major geographic region are noted; bootstrap values above 700 out of 1000 are indicated by thicker lines and bootstrap number.

Fig. 5. Geographic and genetic structure of populations within Africa. (**A**) Geographic discontinuities among African populations using TESS, assuming a model of no population admixture (*25*). Circles indicate location of populations included in the study. (**B**) Inferred proportions of ancestral clusters from STRUCTURE analysis at $K = 14$ for individuals grouped by geographic region and language classification. Classifications of languages spoken by self-identified ethnic affiliation in the Africans are as in Fig. 1. (**C**) Inferred proportion of ancestral clusters in individuals from STRUCTURE analysis at $K = 14$.

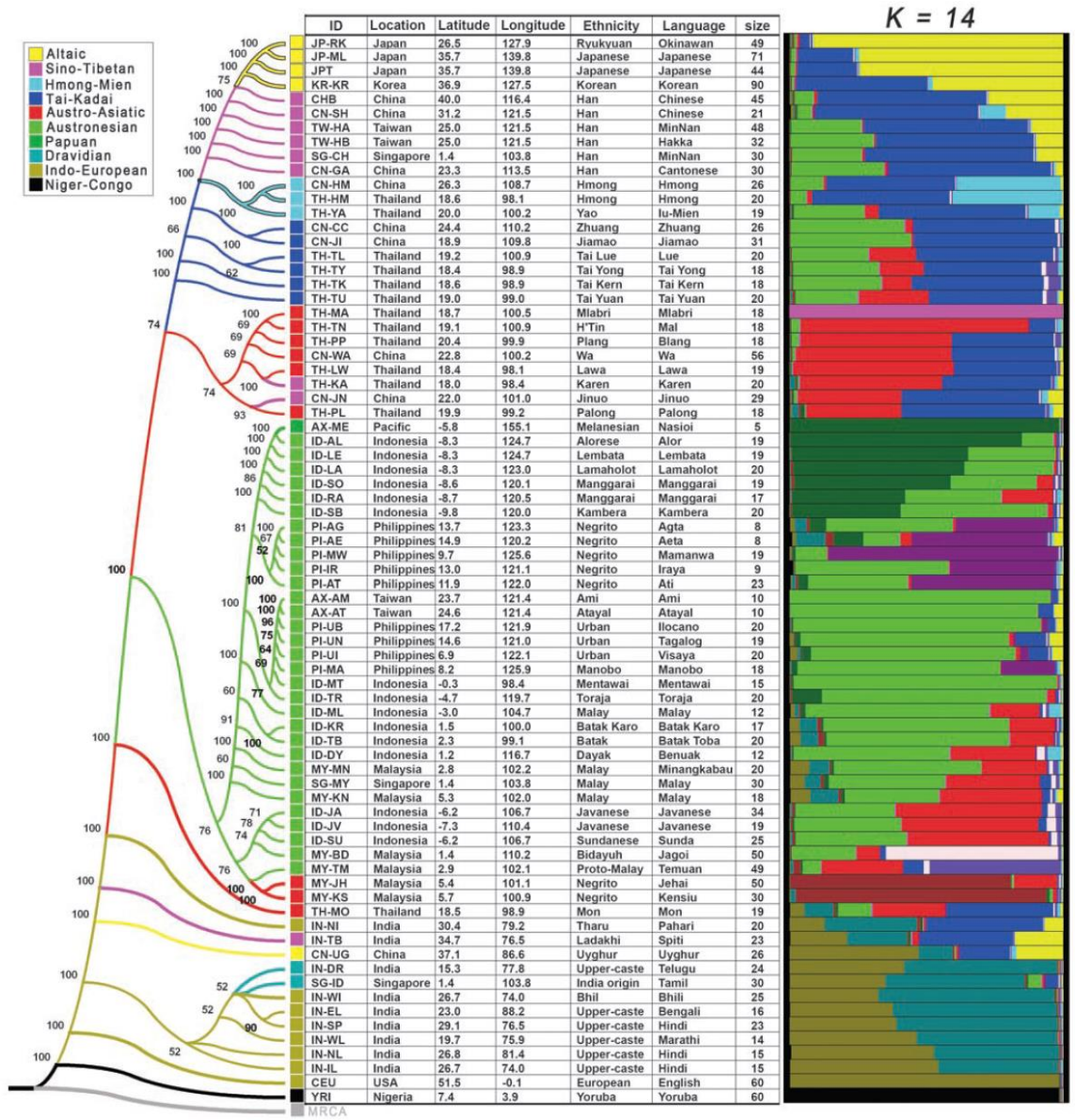


Mapping Human Genetic Diversity in Asia

The HUGO Pan-Asian SNP Consortium*†

Asia harbors substantial cultural and linguistic diversity, but the geographic structure of genetic variation across the continent remains enigmatic. Here we report a large-scale survey of autosomal variation from a broad geographic sample of Asian human populations. Our results show that genetic ancestry is strongly correlated with linguistic affiliations as well as geography. Most populations show relatedness within ethnic/linguistic groups, despite prevalent gene flow among populations. More than 90% of East Asian (EA) haplotypes could be found in either Southeast Asian (SEA) or Central-South Asian (CSA) populations and show clinal structure with haplotype diversity decreasing from south to north. Furthermore, 50% of EA haplotypes were found in SEA only and 5% were found in CSA only, indicating that SEA was a major geographic source of EA populations.

Several genome-wide studies of human genetic diversity focusing primarily on broad continental relationships, or fine-scale structure in Europe, have been published recently (*1*–*8*). We have extended this approach to Southeast Asian (SEA) and East Asian (EA) populations by using the Affymetrix GeneChip Human Mapping 50K Xba Array. Stringently quality-controlled genotypes were obtained at 54,794 autosomal single-nucleotide polymorphisms (SNPs) in 1928 individuals representing 73 Asian and two non-Asian HapMap populations (*9*). Apart from developing a general description of Asian population structure and its relation to geography, language, and demographic history, we concentrated on uncovering the geographic source(s) of EA and SEA populations.



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60
MRCA						

Fig. 1. Maximum-likelihood tree of 75 populations. A hypothetical most-recent common ancestor (MRCA) composed of ancestral alleles as inferred from the genotypes of one gorilla and 21 chimpanzees was used to root the tree. Branches with bootstrap values less than 50% were condensed. Population identification numbers (IDs), sample collection locations with latitudes and longitudes, ethnicities, language spoken, and size of population samples are shown in the table adjacent to each branch in the tree. Linguistic groups are indicated with colors as shown in the legend. All population IDs except the four HapMap samples are denoted by four characters. The first two letters indicate the country where the samples were collected or (in the case of Affymetrix) genotyped, according to the following convention: AX, Affymetrix; CN, China; ID, Indonesia; IN, India; JP, Japan; KR, Korea; MY, Malaysia; PI, the Philippines; SG, Singapore; TH, Thailand; and TW, Taiwan. The last two letters are unique IDs for the population. To the right of the table, an averaged graph of results from STRUCTURE is shown for $K = 14$.

The genome-wide structure of the Jewish people

Doron M. Behar[1,2]*, Bayazit Yunusbayev[2,3]*, Mait Metspalu[2]*, Ene Metspalu[2], Saharon Rosset[4], Jüri Parik[2], Siiri Rootsi[2], Gyaneshwer Chaubey[2], Ildus Kutuev[2,3], Guennady Yudkovsky[1,5], Elza K. Khusnutdinova[3], Oleg Balanovsky[6], Ornella Semino[7], Luisa Pereira[8,9], David Comas[10], David Gurwitz[11], Batsheva Bonne-Tamir[11], Tudor Parfitt[12], Michael F. Hammer[13], Karl Skorecki[1,5] & Richard Villems[2]

Contemporary Jews comprise an aggregate of ethno-religious communities whose worldwide members identify with each other through various shared religious, historical and cultural traditions[1,2]. Historical evidence suggests common origins in the Middle East, followed by migrations leading to the establishment of communities of Jews in Europe, Africa and Asia, in what is termed the Jewish Diaspora[3–5]. This complex demographic history imposes special challenges in attempting to address the genetic structure of the Jewish people[6]. Although many genetic studies have shed light on Jewish origins and on diseases prevalent among Jewish communities, including studies focusing on uniparentally and biparentally inherited markers[7–16], genome-wide patterns of variation across the vast geographic span of Jewish Diaspora communities and their respective neighbours have yet to be addressed. Here we use high-density bead arrays to genotype individuals from 14 Jewish Diaspora communities and compare these patterns of genome-wide diversity with those from 69 Old World non-Jewish populations, of which 25 have not previously been reported. These samples were carefully chosen to provide comprehensive comparisons between Jewish and non-Jewish populations in the Diaspora, as well as with non-Jewish populations from the Middle East and north Africa. Principal component and structure-like analyses identify previously unrecognized genetic substructure within the Middle East. Most Jewish samples form a remarkably tight subcluster that overlies Druze and Cypriot samples but not samples from other Levantine populations or paired Diaspora host populations. In contrast, Ethiopian Jews (Beta Israel) and Indian Jews (Bene Israel and Cochini) cluster with neighbouring autochthonous populations in Ethiopia and western India, respectively, despite a clear paternal link between the Bene Israel and the Levant. These results cast light on the variegated genetic architecture of the Middle East, and trace the origins of most Jewish Diaspora communities to the Levant.

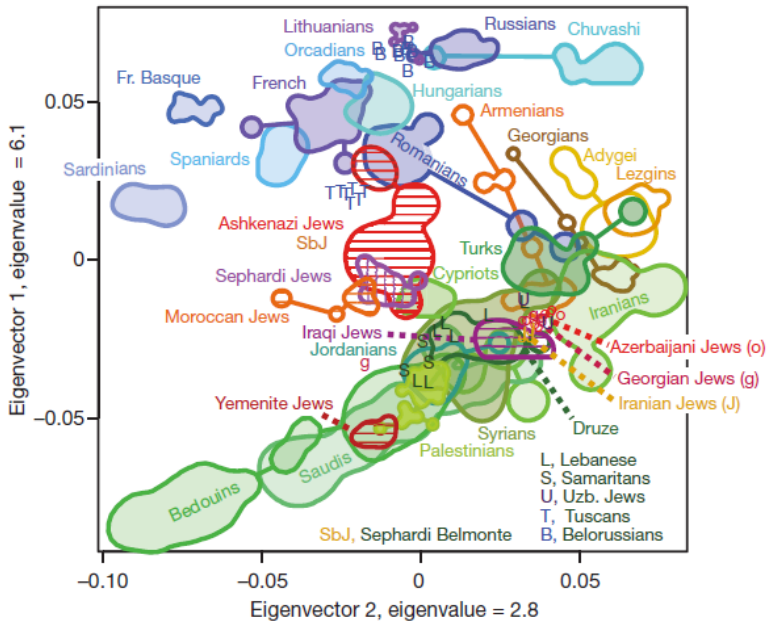


Figure 2 | PCA of west Eurasian high-density array data. Plot of kernel densities (Supplementary Note 2) for each population sample ($n > 10$) was estimated on the basis of PC1 and PC2 coordinates in Supplementary Fig. 3. Individuals from these samples were plotted by using PC1 and PC2 coordinates and were overlaid with the plot of kernel density.

Our PCA, ADMIXTURE and ASD analyses, which are based on genome-wide data from a large sample of Jewish communities, their non-Jewish host populations, and novel samples from the Middle East, are concordant in revealing a close relationship between most contemporary Jews and non-Jewish populations from the Levant. The most parsimonious explanation for these observations is a common genetic origin, which is consistent with an historical formulation of the Jewish people as descending from ancient Hebrew and Israelite residents of the Levant. This inference underscores the significant genetic continuity that exists among most Jewish communities and contemporary non-Jewish Levantine populations, despite their long-term residence in diverse regions remote from the Levant and isolation from one another. This study further uncovers genetic structure that partitions most Jewish samples into Ashkenazi–north African–Sephardi, Caucasus–Middle Eastern, and Yemenite subclusters (Fig. 2). There are several mutually compatible explanations for the observed pattern: a splintering of Jewish populations in the early Diaspora period, an underappreciated level of contact between members of each of these subclusters, and low levels of admixture with Diaspora host populations. Equally interesting are the inferences that can be gleaned from more distant Diaspora communities, such as the Ethiopian and Indian Jewish communities. Strong similarities to their neighbouring host populations may have resulted from one or more of the following: large-scale introgression, asymmetrical sex-biased gene flow, or religious and cultural diffusion during the process of becoming one of the many and varied Jewish communities.

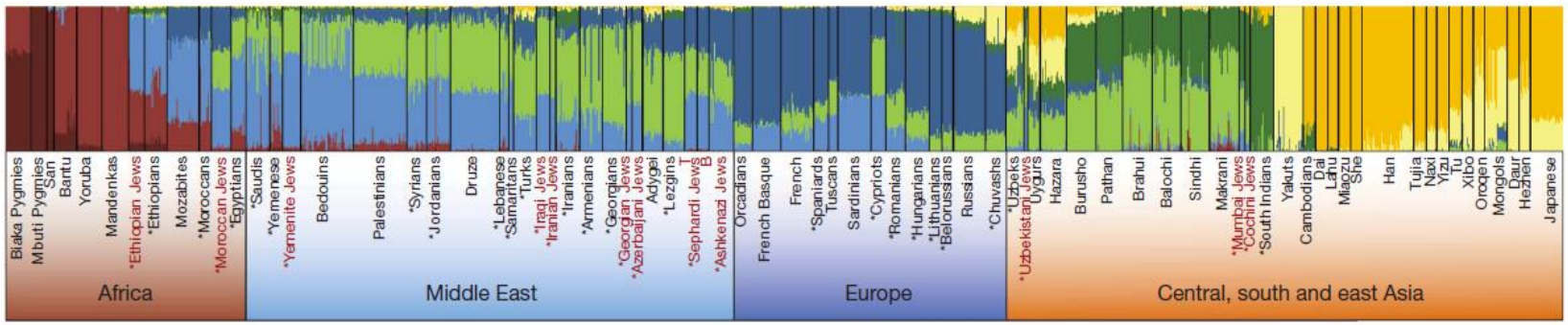


Figure 3 | Population structure inferred by ADMIXTURE analysis. Each individual is represented by a vertical (100%) stacked column of genetic components proportions shown in colour for $K = 8$. The Jewish communities are labelled in colour and bold. T and B further specify Sephardi Jews from Turkey and Bulgaria, respectively. Populations introduced for the first time in this study and analysed together with the Human Genome Diversity Panel[18] data are marked with an asterisk.

Aplikacija genomskih raziskav za humano biologijo

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

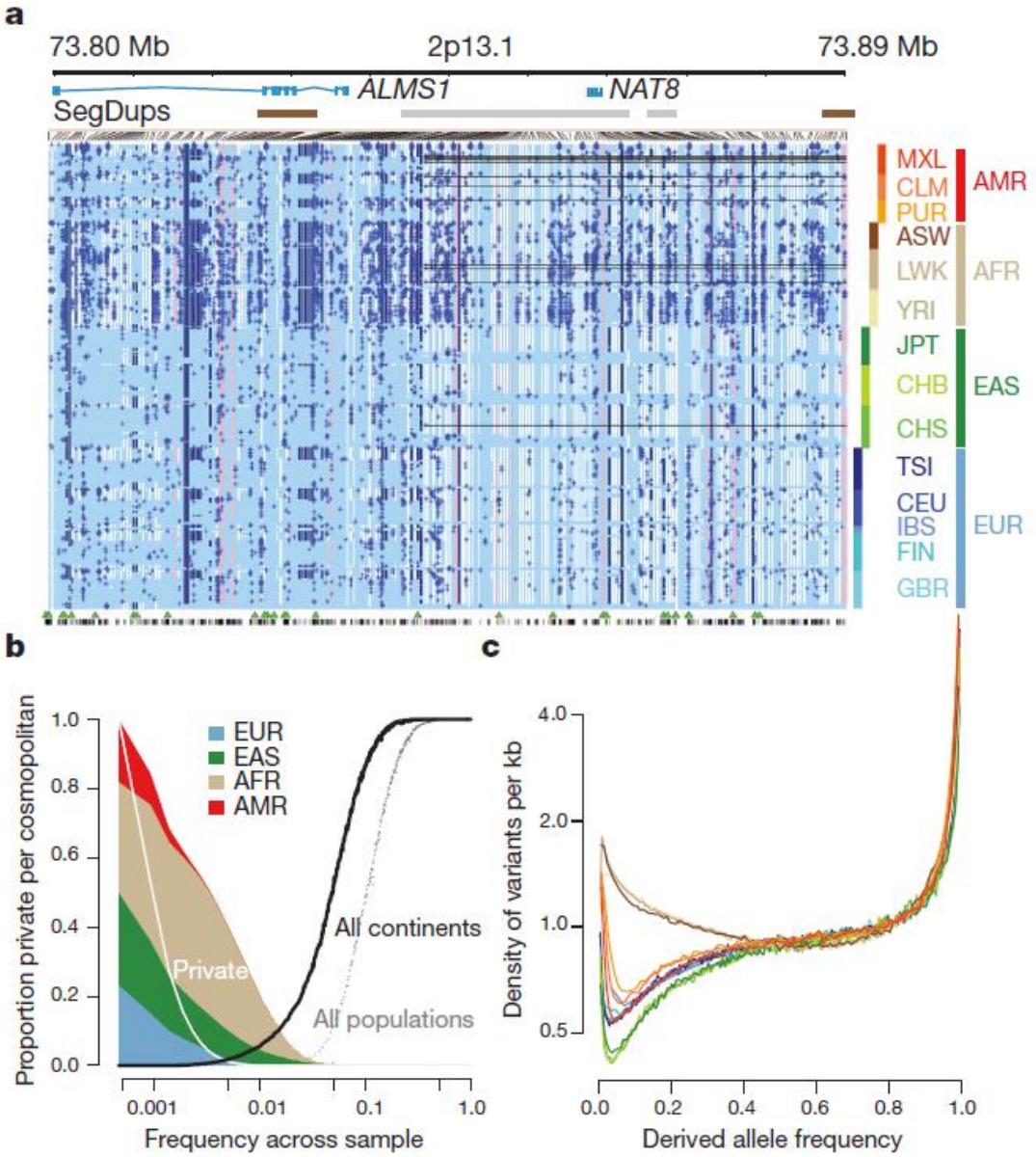


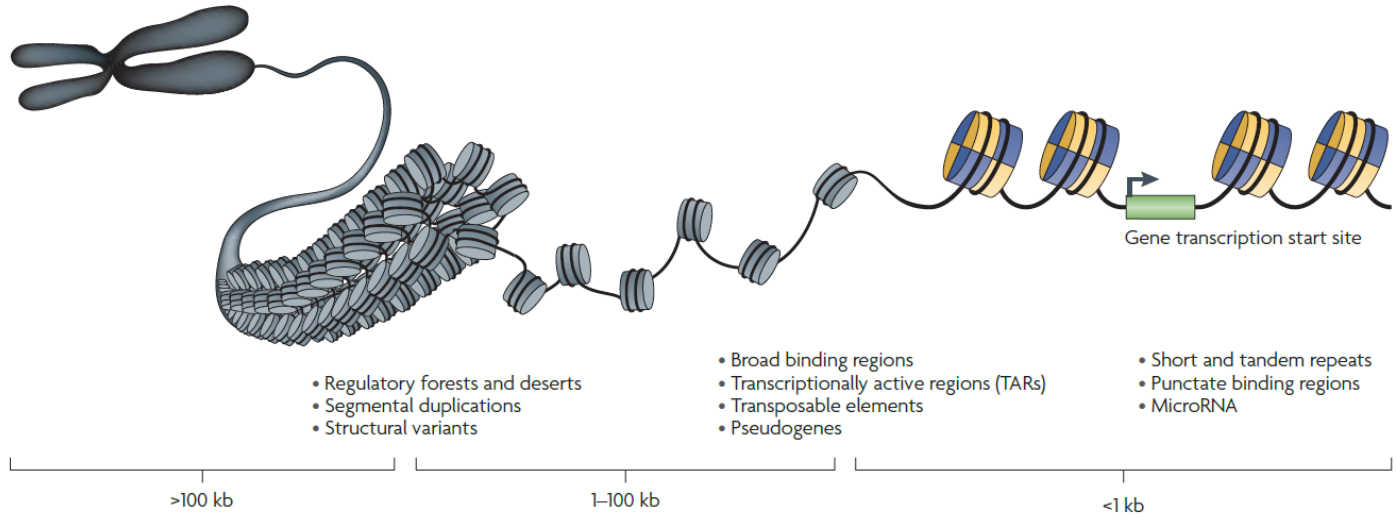
Figure 2 | The distribution of rare and common variants. a, Summary of inferred haplotypes across a 100-kb region of chromosome 2 spanning the genes *ALMS1* and *NAT8*, variation in which has been associated with kidney disease[45].

Annotating non-coding regions of the genome

Roger P. Alexander[*‡], *Gang Fang*[*‡], *Joel Rozowsky*[‡], *Michael Snyder*[§] *and Mark B. Gerstein*[*‡||]

Abstract | Most of the human genome consists of non-protein-coding DNA. Recently, progress has been made in annotating these non-coding regions through the interpretation of functional genomics experiments and comparative sequence analysis. One can conceptualize functional genomics analysis as involving a sequence of steps: turning the output of an experiment into a 'signal' at each base pair of the genome; smoothing this signal and segmenting it into small blocks of initial annotation; and then clustering these small blocks into larger derived annotations and networks. Finally, one can relate functional genomics annotations to conserved units and measures of conservation derived from comparative sequence analysis.

Box 1 | **Catalogue of non-coding elements**



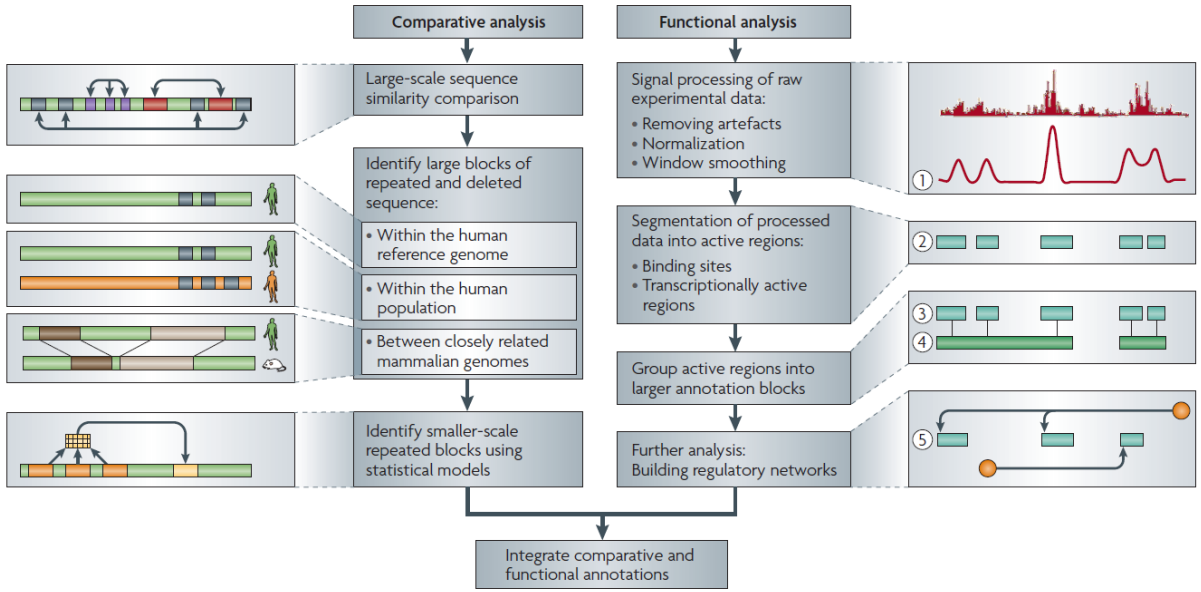


Figure 1 | **Annotation process for non-coding regions: an overview.** The annotation process includes two parallel pipelines for comparative sequence analysis (comparative analysis) and functional genomics analysis (functional analysis) of experimental data. Comparative analysis includes analysis of repeated sequences in the reference human genome, structural variation across the human population and sequence elements conserved across multiple species. The annotation process for functional genomics data involves smoothing the raw signal (step 1), thresholding and segmentation of the smoothed signal (step 2), clustering of discrete segments (step 3), functional annotation of clusters (step 4) and connecting clusters into networks (step 5).

Evolution of genetic and genomic features unique to the human lineage

Majesta O'Bleness[1], Veronica B. Searles[1], Ajit Varki[2,3,4], Pascal Gagneux[3,4] and James M. Sikela[1,4]

Abstract | Given the unprecedented tools that are now available for rapidly comparing genomes, the identification and study of genetic and genomic changes that are unique to our species have accelerated, and we are entering a golden age of human evolutionary genomics. Here we provide an overview of these efforts, highlighting important recent discoveries, examples of the different types of human-specific genomic and genetic changes identified, and salient trends, such as the localization of evolutionary adaptive changes to complex loci that are highly enriched for disease associations. Finally, we discuss the remaining challenges, such as the incomplete nature of current genome sequence assemblies and difficulties in linking human-specific genomic changes to human-specific phenotypic traits.

Table 1 | **Partial list of genes and genetic elements showing human-lineage-specific changes**

Gene or element	Mechanism of change	Proposed phenotype	Phenotypic certainty	Possible gene-associated diseases	Refs
AR	Deletion of regulatory DNA	Loss of sensory vibrissae and penile spines	Likely	Androgen insensitivity; hypospadias; muscular atrophy; prostate cancer	1
APOC1	Pseudogene	Unknown	Not applicable	Alzheimer's severity; atherosclerosis; coronary heart disease	55–59
AQP7	Copy number increase	Energy use	Plausible	Nonfunctional glycerol response to exercise	10,73–75
ASPM	Positive selection	Increased brain size	Plausible	Microcephaly	94,95
CDK5RAP2	Positive selection	Increased brain size	Plausible	Microcephaly	95,118
CCL3L1	Novel gene variant	Immune system function	Likely	HIV and AIDS; Kawasaki's disease; rheumatoid arthritis; chronic hepatitis C	89
CHRM3	Novel exon	Change in human reproduction	Plausible	Eagle–Barrett syndrome	38
CHRFAM7A	Copy number increase	Higher brain function	Plausible	P50 sensory gating deficit	8,89,119
CMAH	Pseudogene	Changed sialic acid composition on all cells	Definite	Duchenne's muscular dystrophy; red-meat-related carcinoma risk	62,63
COX5A	Amino acid change	Mitochondrial metabolism	Plausible	Unknown	49
DRD5	Copy number increase	Regulation of memory; attention; movement	Likely	DRD5 deficiency; attention-deficit hyperactivity disorder; primary cervical dystonia	8,89
DUF1220 and NBPF family	Protein domain copy number increase	Brain size	Likely	Microcephaly; macrocephaly	41,42, 45,46
FCGR1A	Copy number increase	Immune system function	Plausible	IgG receptor I phagocyte deficiency	25,89
FSHR	Positive selection	Decreased gestation; birth timing	Plausible	Amenorrhoea; infertility; ovarian dysgenesis type 1; ovarian hyperstimulation syndrome	120,121
FOXP2	Amino acid change	Speech and language development	Definite	Speech and language disorder 1	51

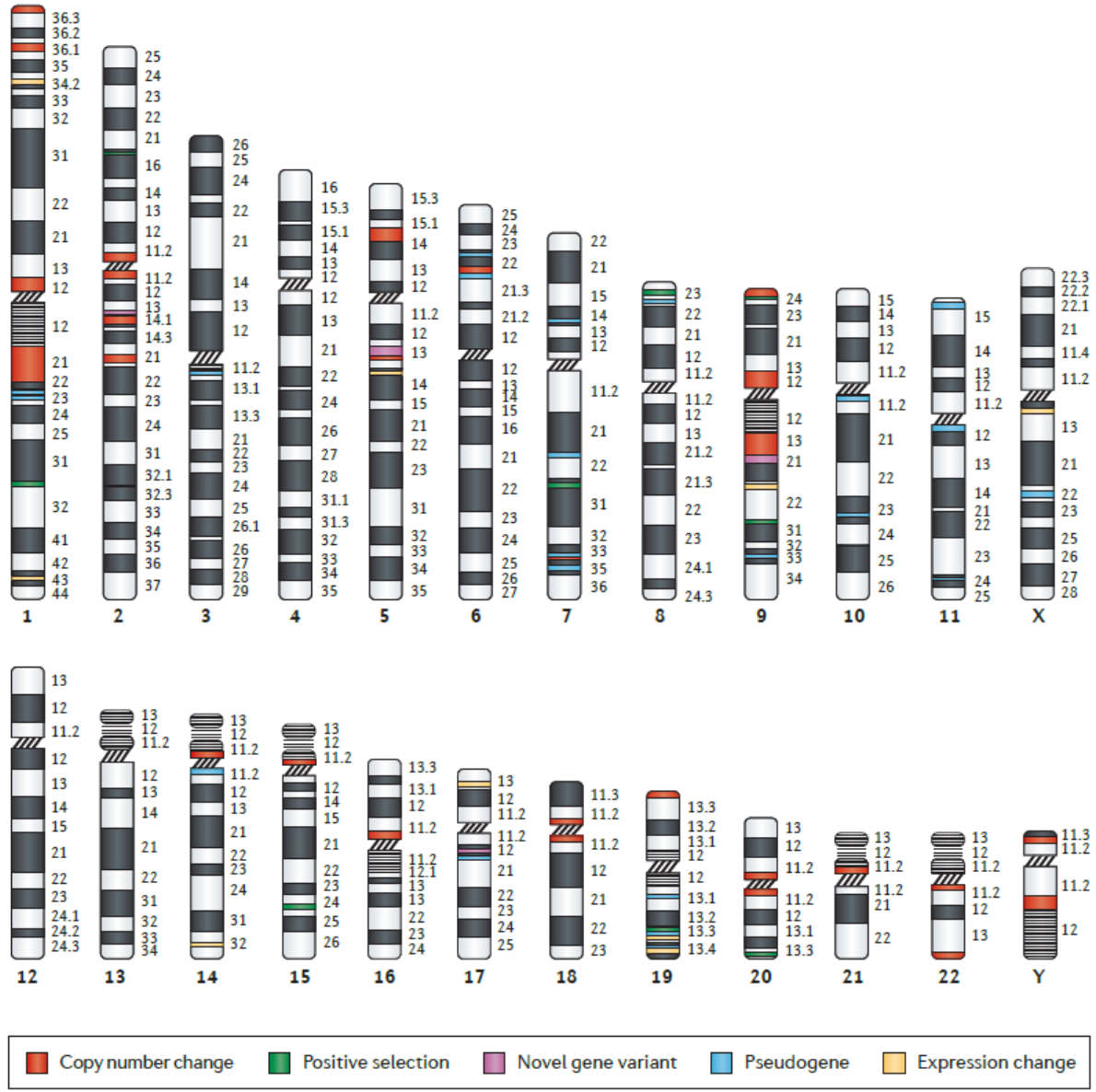
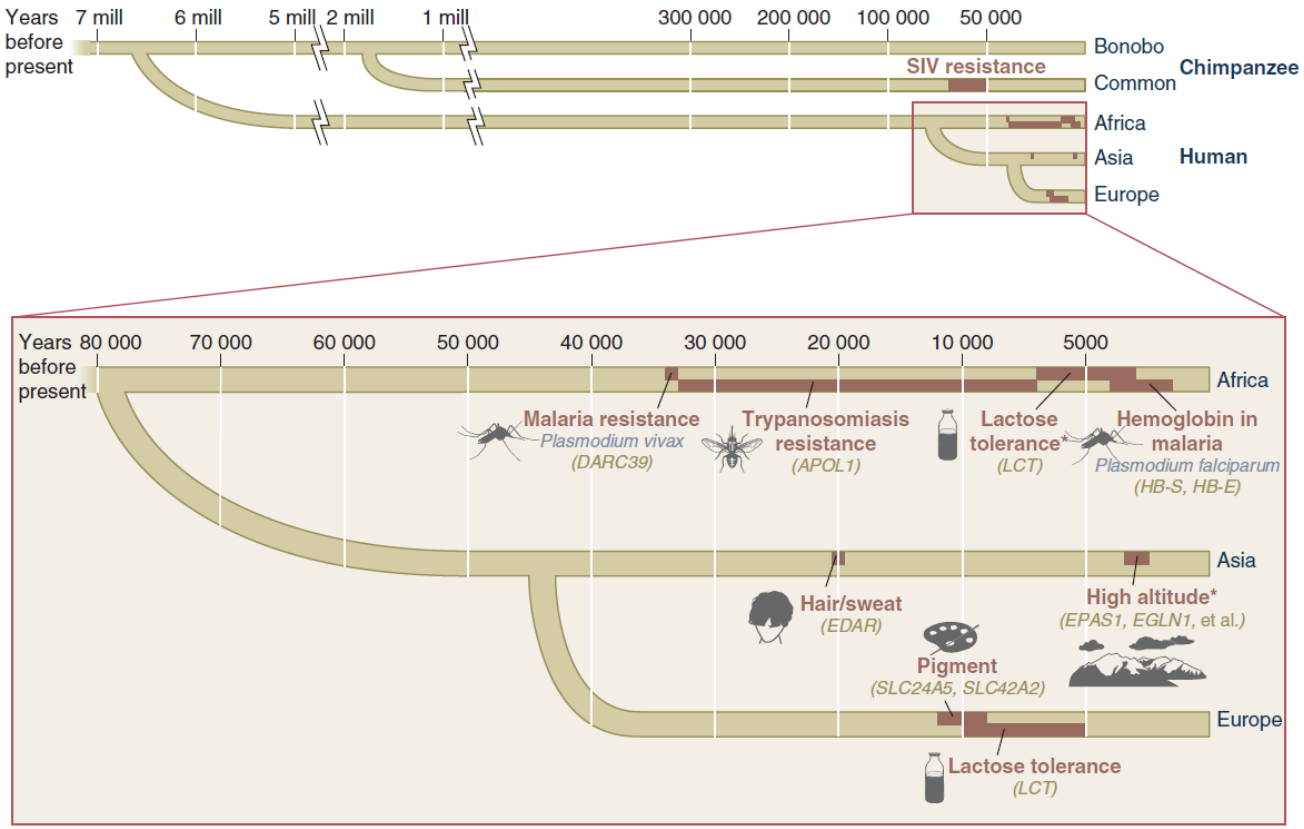


Figure 1 | **Genome positions of human-lineage-specific gene changes.** Human-lineage-specific (HLS) gene changes discussed in this paper are displayed in their corresponding genomic position across the human karyotype. The changes are divided into five categories that correspond to those listed in TABLE 1, and each type is colour coded. It should be noted that many genes have undergone multiple types of HLS changes, and in this case only one type is shown. For visualization purposes, the size of the coloured bands is not drawn to scale.

Recent human adaptation: genomic approaches, interpretation and insights

Laura B. Scheinfeldt[1,2] and Sarah A. Tishkoff[1,3]

Abstract | The recent availability of genomic data has spurred many genome-wide studies of human adaptation in different populations worldwide. Such studies have provided insights into novel candidate genes and pathways that are putatively involved in adaptation to different environments, diets and disease prevalence. However, much work is needed to translate these results into candidate adaptive variants that are biologically interpretable. In this Review, we discuss methods that may help to identify true biological signals of selection and studies that incorporate complementary phenotypic and functional data. We conclude with recommendations for future studies that focus on opportunities to use integrative genomics methodologies in human adaptation studies.



TRENDS in Genetics

Evidence from the archaeological record as well as from genetic and genomic studies demonstrates that anatomically modern humans emerged in Africa ~200 thousand years ago (kya) and rapidly migrated across the globe into new environments ~80–50 kya (reviewed in REFS 1,2). One of the implications of this history is that our Late Pleistocene (~125–12 kya) ancestors were subjected to a diverse range of novel selective pressures. Therefore, phenotypes such as thermoregulation in cold environments, tolerance to hypoxia at high altitude and light skin pigmentation in regions with low amounts of sunlight are likely to have increased reproductive fitness and thus to have been affected by adaptive pressures. In addition, during the Neolithic period (~12–4 kya), many of our ancestors began to adopt more sedentary lifestyles, resulting in increased population densities, as well as distinct diets that were associated with pastoral and agricultural technologies. These changes are also likely to have resulted in distinct adaptive pressures. For example, increased population densities are correlated with increased infectious-disease loads, and phenotypes related to the immune response are thought to have been affected by this environmental shift[3]. Thus, the identification of genetic signatures of human adaptation to such pressures is informative not only for understanding the ways in which adaptation has shaped genetic variation in contemporary populations, but also because the phenotypic consequences of adaptive genetic variants may have a role in the biological variation and health of contemporary humans.

How culture shaped the human genome: bringing genetics and the human sciences together

Kevin N. Laland, John Odling-Smee‡ and Sean Myles§ ||*

Abstract | Researchers from diverse backgrounds are converging on the view that human evolution has been shaped by gene–culture interactions. Theoretical biologists have used population genetic models to demonstrate that cultural processes can have a profound effect on human evolution, and anthropologists are investigating cultural practices that modify current selection. These findings are supported by recent analyses of human genetic variation, which reveal that hundreds of genes have been subject to recent positive selection, often in response to human activities. Here, we collate these data, highlighting the considerable potential for cross-disciplinary exchange to provide novel insights into how culture has shaped the human genome.

Box 5 | **Genetic responses to human diet**

Cultural variation in human diet clearly explains some of the adaptive genetic differences between human populations. One compelling example of a human-culture-initiated selective sweep concerns the evolution of the human amylase gene[112]. Starch consumption is a feature of agricultural societies and hunter-gatherers in arid environments, whereas other hunter-gatherers and some pastoralists consume much less starch. This behavioural variation raises the possibility that different selective pressures have acted on amylase, the enzyme responsible for starch hydrolysis. Consistent with this hypothesis, Perry *et al.*[112] found that copy number of the salivary amylase gene (*AMY1*) is positively correlated with salivary amylase protein level and that individuals from populations with high-starch diets have, on average, more *AMY1* copies than those with traditionally low-starch diets. Higher *AMY1* copy numbers and protein levels are thought to improve the digestion of starchy foods and may buffer against the fitness-reducing effects of intestinal disease.

More generally, the transition to novel food sources with the advent of agriculture and the colonization of new habitats seems to have been a major source of selection on human genes[6,113]. Wang *et al.*[7] describe protein metabolism as an overrepresented category (15%) in selective events, and affected genes include ADAM metallopeptidase with thrombospondin motif 19 (*ADAMTS19*), *ADAMTS20*, *N*-acylaminoacyl-peptide hydrolase (*APEH*), plasminogen activator, urokinase (*PLAU*), histone deacetylase 8 (*HDAC8*), ubiquitin protein ligase E3 component n-recognin 1 (*UBR1*) and ubiquitin-specific peptidase 26 (*USP26*). Several genes related to the metabolism of carbohydrates, lipids and phosphates also show signals of recent selection, including genes involved in metabolizing mannose (*MAN2A1* in Yorubans and East Asians), sucrose (*SI* in East Asians) and fatty acids (solute carrier family 27, member 4 (*SLC27A4*) and peroxisome proliferator-activated receptor δ (*PPARD*) in Europeans, *SLC25A20* in East Asians, nuclear receptor coactivator 1 (*NCOA1*) in Yorubans and leptin receptor (*LEPR*) in East Asians)[6]. Williamson *et al.*[14] add sterol carrier protein 2 (*SCP2*), which has a role in the intracellular movement of cholesterol. There is also evidence for diet-related selection on the thickness of human teeth enamel[114] and bitter-taste receptors[115], and the promoter regions of many nutrition-related genes have experienced positive selection during human evolution[116]. In addition, there is a strong signal of selection in the alcohol dehydrogenase (*ADH*) cluster in East Asians, which is thought to be an interaction between the functional polymorphisms of the alcohol-metabolism genes in protection against alcoholism[117]. One argument is that hypersensitivity to alcohol has been adaptive through protecting against alcoholism[118].

Table 2 | **Genes identified as having been subject to recent rapid selection and their inferred cultural selection pressures**

Genes	Function or phenotype	Inferred cultural selection pressure	Refs
LCT, MAN2A1, SI, SLC27A4, PPARD, SLC25A20, NCOA1, LEPR, LEPR, ADAMTS19, ADAMTS20, APEH, PLAU, HDAC8, UBR1, USP26, SCP2, NKX2-2, AMY1, ADH, NPY1R, NPY5R	Digestion of milk and dairy products; metabolism of carbohydrates, starch, proteins, lipids and phosphates; alcohol metabolism	Dairy farming and milk usage; dietary preferences; alcohol consumption	6,7,16,41,63, 102,118, 144,145
Cytochrome P450 genes (*CYP3A5, CYP2E1, CYP1A2* and *CYP2D6*)	Detoxification of plant secondary compounds	Domestication of plants	6,63,146,147
CD58, APOBEC3F, CD72, FCRL2, TSLP, RAG1, RAG2, CD226, IGJ, TJP1, VPS37C, CSF2, CCNT2, DEFB118, STAB1, SP1, ZAP70, BIRC6, CUGBP1, DLG3, HMGCR, STS, XRN2, ATRN, G6PD, TNFSF5, HbC, HbE, HbS, Duffy, α-globin	Immunity, pathogen response; resistance to malaria and other crowd diseases	Dispersal, agriculture, aggregation and subsequent exposure to new pathogens; farming	6–8,14,16,50, 63,148,149
LEPR, PON1, RAPTOR, MAPK14, CD36, DSCR1, FABP2, SOD1, CETP, EGFR, NPPA, EPHX2, MAPK1, UCP3, LPA, MMRN1	Energy metabolism, hot or cold tolerance; heat-shock genes	Dispersal and subsequent exposure to novel climates	14,150
SLC24A5, SLC25A2, EDAR, EDA2R, SLC24A4, KITLG, TYR, 6p25.3, OCA2, MC1R, MYO5A, DTNBP1, TYRP1, RAB27A, MATP, MC2R, ATRN, TRPM1, SILV, KRTAPs, DCT	The externally visible phenotype (skin pigmentation, hair thickness, eye and hair colour, and freckles)	Dispersal and local adaptation and/or sexual selection	9,14,63,97, 101,151
CDK5RAP2, CENPJ, GABRA4, PSEN1, SYT1, SLC6A4, SNTG1, GRM3, GRM1, GLRA2, OR4C13, OR2B6, RAPSN, ASPM, RNT1, SV2B, SKP1A, DAB1, APPBP2, APBA2, PCDH15, PHACTR1, ALG10, PREP, GPM6A, DGKI, ASPM, MCPH1, FOXP2	Nervous system, brain function and development; language skills and vocal learning	Complex cognition on which culture is reliant; social intelligence; language use and vocal learning	6,7,14,63, 68–70,78,149
BMP3, BMPR2, BMP5, GDF5	Skeletal development	Dispersal and sexual selection	6,63
MYH16, ENAM	Jaw muscle fibres; tooth-enamel thickness	Invention of cooking; diet	80,113

There are reasons to anticipate that gene–culture interactions may have had a prominent role in local, geographically restricted adaptation over the past 50,000 years. Not only did humans recently come to occupy nearly every habitable corner of the earth, but cultures rapidly diversified during the out-of-Africa expansion. Moreover, human culture is cumulative, with tools and technology building on earlier forms, which implies that humans must possess more culture, and more potent culture, now than earlier in history. For comparison, the lithic technology of early *Homo* species remained largely unchanged for a million years. These considerations imply an increasing significance of gene–culture co-evolution with time. A gene–culture co-evolutionary perspective predicts that the genetic signatures of recent positive selection (for example, since the out-of-Africa expansion) will more often have been generated by culture than signatures of selection from earlier time periods in human evolution (for example, before the out-of-Africa expansion).

Genomic Data Reveal a Complex Making of Humans

Isabel Alves[1,2,3☯], **Anna Šrámková Hanulová**[1,2☯], **Matthieu Foll**[1,2], **Laurent Excoffier**[1,2*]

1 CMPG, Institute of Ecology and Evolution, Berne, Switzerland, **2** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **3** Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, Oeiras, Portugal

Abstract: In the last few years, two paradigms underlying human evolution have crumbled. Modern humans have not totally replaced previous hominins without any admixture, and the expected signatures of adaptations to new environments are surprisingly lacking at the genomic level. Here we review current evidence about archaic admixture and lack of strong selective sweeps in humans. We underline the need to properly model differential admixture in various populations to correctly reconstruct past demography. We also stress the importance of taking into account the spatial dimension of human evolution, which proceeded by a series of range expansions that could have promoted both the introgression of archaic genes and background selection.

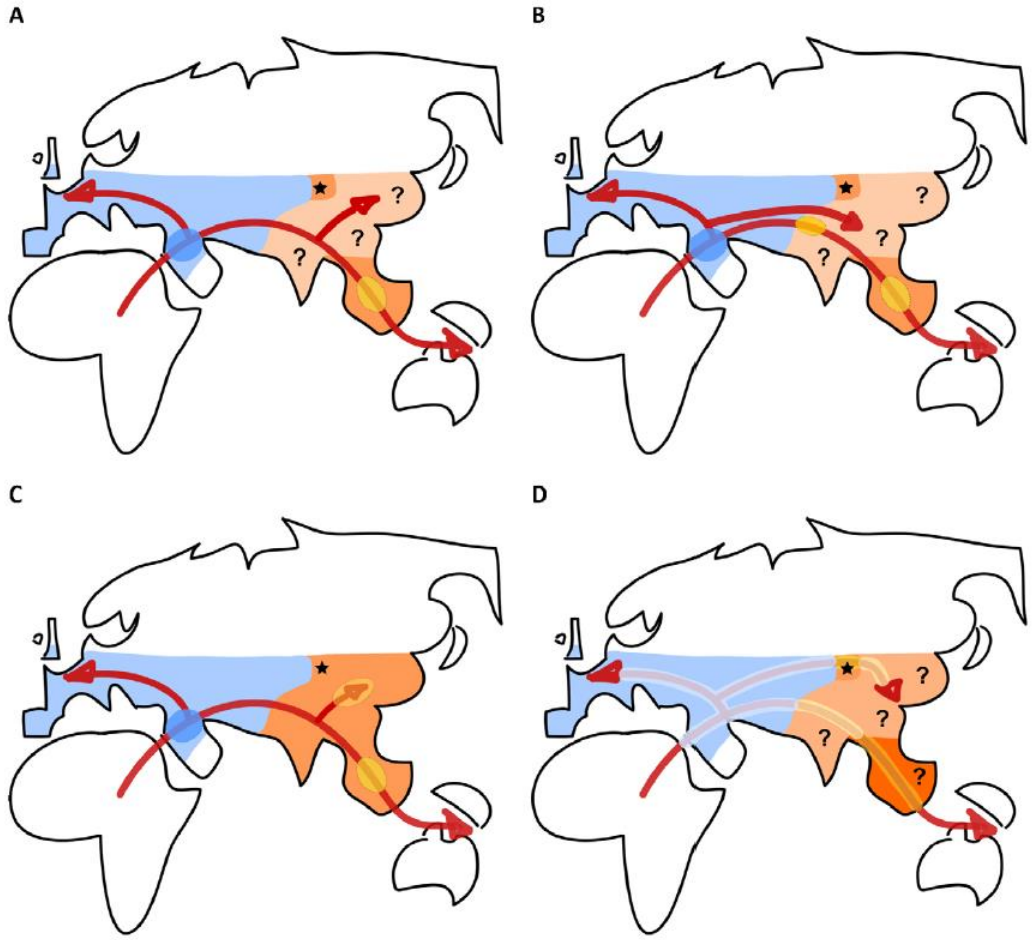


Figure 1. Sketches of different scenarios of human dispersal and admixture with archaic human populations during their range expansion out of Africa. Red arrows indicate approximate migration routes. Neanderthal range is in blue, Denisovan range(s) in orange, and the location of the Denisova site is indicated as a black star. Question marks in the Denisovan range indicate uncertainty on Denisovan hominin presence. Filled ellipses indicate potential places of admixture in scenarios (A–C). (A) Scenario of Reich et al. [15,16] with pulses of admixture between modern humans and Neanderthals (dark blue ellipse) and between modern humans and Denisovans (yellow ellipse). (B) Scenario of Rasmussen et al. [24] with two waves into Asia. Denisovan admixture in Oceanians would have occurred during the first wave, possibly at different places during the migration. (C) Scenario of Skoglund and Jakobsson [17], with distinct Denisovan admixture events in Oceanians and East Asians. (D) Extension of the spatially explicit scenario of Currat and Excoffier [25] postulating a continuous admixture between modern humans and archaic hominins along migration routes overlapping with archaic hominin ranges. Different shades of orange indicate potentially different archaic hominin populations in Asia.
doi:10.1371/journal.pgen.1002837.g001

Learning about human population history from ancient and modern genomes

Mark Stoneking and Johannes Krause‡*

Abstract | Genome-wide data, both from SNP arrays and from complete genome sequencing, are becoming increasingly abundant and are now even available from extinct hominins. These data are providing new insights into population history; in particular, when combined with model-based analytical approaches, genome-wide data allow direct testing of hypotheses about population history. For example, genome-wide data from both contemporary populations and extinct hominins strongly support a single dispersal of modern humans from Africa, followed by two archaic admixture events: one with Neanderthals somewhere outside Africa and a second with Denisovans that (so far) has only been detected in New Guinea. These new developments promise to reveal new stories about human population history, without having to resort to storytelling.

New methods for inferring history from genome-wide data. Human population history may be about telling stories, but we need better ways to discern what happened in the past without resorting to storytelling. Substantial advances have been made in applying model-based approaches for testing different models of population history and for estimating demographic parameters corresponding to the best-fitting model[52,53]. Also, although it is still challenging, some progress has been made in methods for estimating the timing of old admixture events and investigating admixture involving closely related populations[59,60]. However, there is much more that needs to be done, particularly in fitting more complex (and hence realistic) models.

The good news (especially for students who may fear that all interesting questions have been answered) is that when it comes to human population history, there are still many stories waiting to be told.

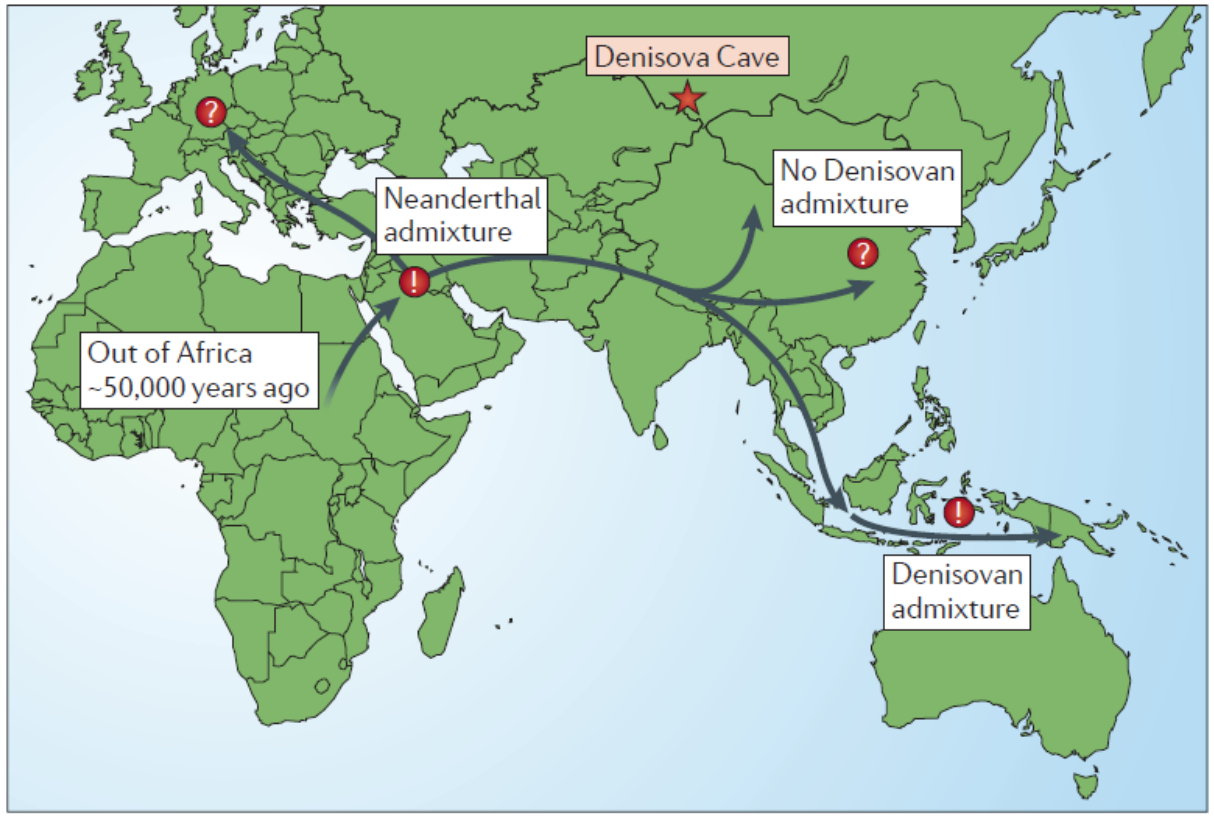


Figure 4 | **Dispersal of modern humans from Africa.** A map illustrating the dispersal of modern humans from Africa about 50,000 years ago, followed by admixture with Neanderthals in the ancestry of all non-Africans, followed by admixture with Denisovans in the ancestry of New Guineans. Arrows indicate general directionality and not specific migration routes — in general we only know for sure the end points of migrations, not the routes. The red star indicates the location of Denisova Cave. The exclamation marks indicate admixture, but there is extreme uncertainty as to where the Neanderthal and Denisovan admixture occurred. Question marks indicate regions where no additional admixture was detected even though archaeological findings suggest that Neanderthals and Denisovans overlapped with modern human populations in those regions.

Mapping Human Epigenomes

Chloe M. Rivera[1,2] and Bing Ren[1,3,*]
[1]Ludwig Institute for Cancer Research
[2]The Biomedical Sciences Graduate Program
[3]Department of Cellular and Molecular Medicine
Institute of Genomic Medicine, UCSD Moores Cancer Center, University of California School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0653, USA
*Correspondence: biren@ucsd.edu
http://dx.doi.org/10.1016/j.cell.2013.09.011

As the second dimension to the genome, the epigenome contains key information specific to every type of cells. Thousands of human epigenome maps have been produced in recent years thanks to rapid development of high throughput epigenome mapping technologies. In this review, we discuss the current epigenome mapping toolkit and utilities of epigenome maps. We focus particularly on mapping of DNA methylation, chromatin modification state, and chromatin structures, and emphasize the use of epigenome maps to delineate human gene regulatory sequences and developmental programs. We also provide a perspective on the progress of the epigenomics field and challenges ahead.

Table 1. Large-Scale National and International Epigenomic Consortia

Project Name	Start Date	Affiliations	Completed and Expected Data Contributions	Selected Publication	Access Data
Encyclopedia of DNA Elements	2003	NIH	Dnase-seq, RNA-seq, ChIP-seq, and 5C in 100s of primary human tissues and cell lines	ENCODE Project Consortium et al., 2012	http://encodeproject.org/ENCODE/
The Cancer Genome Atlas (TCGA)	2006	NIH	DNA methylomes in 1,000s of patients samples from more than 20 cancer types	Garraway and Lander, 2013	http://cancergenome.nih.gov/
Roadmap Epigenomics Project	2008	NIH	Dnase-seq, RNA-seq, ChIP-seq, and MethylC-seq in 100 s of normal primary cells, hESC, and hESC derived cells	Bernstein et al., 2010	http://www.epigenomebrowser.org/
International Cancer Genome Consortium (ICGC)	2008	15 countries, includes TCGA	DNA methylation profiles in thousands of patient samples from 50 different cancers	The International Cancer Genome Consortium, et al., 2010	http://dcc.icgc.org/web
International Human Epigenome Consortium (IHEC)	2010	7 countries, includes BLUEPRINT, Roadmap	Goal: 1,000 Epigenomes in 250 cell types	American Association for Cancer Research Human Epigenome Task Force; European Union, Network of Excellence, Scientific Advisory Board, 2008	http://ihec-epigenomes.org

Table 2. Major Conceptual Advances Convey the Utility of Epigenome Maps

Before Next-Gen Sequencing	The Next-Gen Sequencing Era	Future[a]
DNA Methylation		
Repressive mark at imprinted loci, transposons, and in x chromosome inactivation	Metastable mark	Comprehensive definition of epigenomic variation across all human cell types
Found in active gene bodies	Active DNA demethylation by TET family proteins through 5hmC, 5fC, and 5caC	Define epigenomic variation in populations
Only dynamic in primordial germ cells and during early embryogenesis; in CpG context	Non-CpG methylation exists especially in ESCs oocytes, and adult brain	Discovering epigenetic signatures of disease
	Tissue-specific methylation at distal regulatory elements	Technology advances to enable single-cell methylomes
Histone Modifications		A deeper understanding of epigenomic inheritabilty and reprogramming
Marks that correlate with promoters and gene bodies	Over 130 different histone modifications have been identified	High-throughput functional validation of predicted enhancers
H3K9me3 is a mark of heterochromatin	dentification of novel ncRNAs by promoter and gene body chromatin signatures	Epigenome engineering
H3K27me3 is a mark of facultative heterochromatin	Active enhancers are marked by H3K27ac or H4K16ac	Improved computational analysis and visualization tools
Proposal of the Histone Code Hypothesis	Poised enhancers are marked by H3K4me1 alone or in combination with H3K27me3	
Bivalent Promoters marked by H3K4me3/ H3K27me3	Expansion of repressive chromatin blocks during differentiation	
Unique chromatin signature of enhancers defined as H3K4me1	Combinations of chromatin marks define a limited number of chromatin states	
Chromatin Structure		
Nucleosome maps only in yeast, fly	Nucleosomes mapped in the human genome	
DHSs correlate with TF-binding sites and regulatory elements	Nucleosomes are well positioned around regulatory regions	
	DHSs predict enhancer-promoter pairs and cell types affected in disease	
Nuclear Architecture		
Identification of chromosome territories, LADS, and transcription factories with FISH	Identification of sub-TADs, TADs and chromosome compartments	
Few validated enhancer-promoter interacting pairs	Chromosome-wide maps of enhancer, promoter, and insulator interacting pairs	

[a]Text in this column refers to DNA Methylation, Histone Modifications, Chromatin Structure, and Nuclear Architecture.

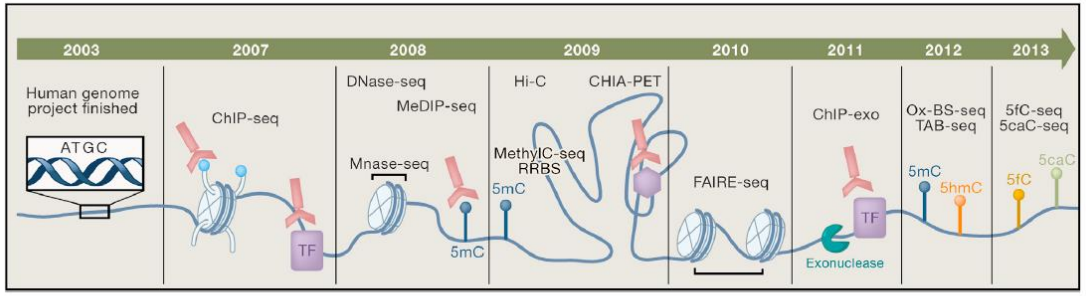


Figure 1. Timeline of Sequencing-Based Technologies for Mapping Human Epigenomes

Aplikacija genomskih raziskav za biologijo

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2]*, Michael F. Lin[1,2]*, Pouya Kheradpour[2]*, Jakob S. Pedersen[3,4]*, Leopold Parts[5,6], Joseph W. Carlson[7], Madeline A. Crosby[8], Matthew D. Rasmussen[2], Sushmita Roy[9], Ameya N. Deoras[2], J. Graham Ruby[10,11], Julius Brennecke[12], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†, Emily Hodges[12], Angie S. Hinrichs[4], Anat Caspi[13], Benedict Paten[4,5,14], Seung-Won Park[15], Mira V. Han[16], Morgan L. Maeder[17], Benjamin J. Polansky[17], Bryanne E. Robson[17], Stein Aerts[18,19], Jacques van Helden[20], Bassem Hassan[18,19], Donald G. Gilbert[21], Deborah A. Eastman[17], Michael Rice[22], Michael Weir[23], Matthew W. Hahn[16], Yongkyu Park[15], Colin N. Dewey[24], Lior Pachter[25,26], W. James Kent[4], David Haussler[4], Eric C. Lai[27], David P. Bartel[10,11], Gregory J. Hannon[12], Thomas C. Kaufman[21], Michael B. Eisen[28,29], Andrew G. Clark[30], Douglas Smith[31], Susan E. Celniker[7], William M. Gelbart[8,32] & Manolis Kellis[1,2]

Sequencing of multiple related species followed by comparative genomics analysis constitutes a powerful approach for the systematic understanding of any genome. Here, we use the genomes of 12 *Drosophila* species for the *de novo* discovery of functional elements in the fly. Each type of functional element shows characteristic patterns of change, or 'evolutionary signatures', dictated by its precise selective constraints. Such signatures enable recognition of new protein-coding genes and exons, spurious and incorrect gene annotations, and numerous unusual gene structures, including abundant stop-codon readthrough. Similarly, we predict non-protein-coding RNA genes and structures, and new microRNA (miRNA) genes. We provide evidence of miRNA processing and functionality from both hairpin arms and both DNA strands. We identify several classes of pre- and post-transcriptional regulatory motifs, and predict individual motif instances with high confidence. We also study how discovery power scales with the divergence and number of species compared, and we provide general guidelines for comparative studies.

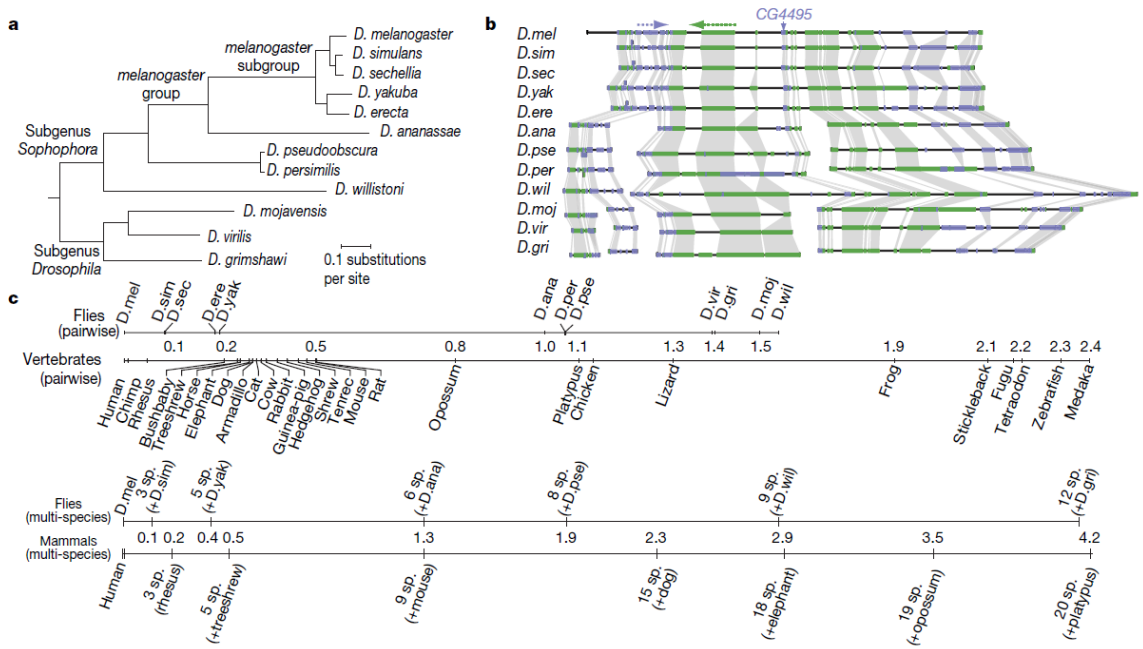


Figure 1 | Phylogeny and alignment of 12 *Drosophila* species. **a**, Phylogenetic tree relating the 12 *Drosophila* species, estimated from fourfold degenerate sites (Supplementary Methods 1). The 12 species span a total branch length of 4.13 substitutions per neutral site. **b**, Gene order conservation for a 0.45-Mb region of chromosome 2L centred on *CG4495*, for which we predict a new exon (Fig. 3a), and spanning 35 genes. Colour represents the direction of transcription. Boxes represent full gene models. Individual exons and introns are not shown. **c**, Comparison of evolutionary distances spanned by fly and vertebrate trees. Pairwise and multi-species distances (in substitutions per fourfold degenerate site) are shown from *D. melanogaster* and from human as reference genomes. Note that species with longer branches (for example, mouse) show higher pairwise distances, not always reflecting the order of divergence. Multi-species distances include all species within a phylogenetic clade.

Evolution of genes and genomes on the *Drosophila* phylogeny

Drosophila 12 Genomes Consortium*

Comparative analysis of multiple genomes in a phylogenetic framework dramatically improves the precision and sensitivity of evolutionary inference, producing more robust results than single-genome analyses can provide. The genomes of 12 *Drosophila* species, ten of which are presented here for the first time (*sechellia*, *simulans*, *yakuba*, *erecta*, *ananassae*, *persimilis*, *willistoni*, *mojavensis*, *virilis* and *grimshawi*), illustrate how rates and patterns of sequence divergence across taxa can illuminate evolutionary processes on a genomic scale. These genome sequences augment the formidable genetic tools that have made *Drosophila melanogaster* a pre-eminent model for animal genetics, and will further catalyse fundamental research on mechanisms of development, cell biology, genetics, disease, neurobiology, behaviour, physiology and evolution. Despite remarkable similarities among these *Drosophila* species, we identified many putatively non-neutral changes in protein-coding genes, non-coding RNA genes, and *cis*-regulatory regions. These may prove to underlie differences in the ecology and behaviour of these diverse species.

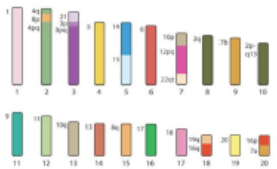
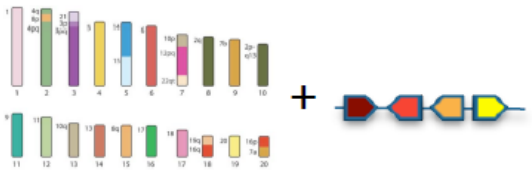
Virtually any question about the function of genome features in *Drosophila* is now empowered by being embedded in the context of this 12 species phylogeny, allowing an analysis of the ways by which evolution has tuned myriad biological processes across the hundreds of millions of years spanned in total by this phylogeny. The analyses presented herein have generated more questions than they have answered, and these results represent a small fraction of that which is possible. Because much of this rich and extraordinary comparative genomic dataset remains to be explored, we believe that these 12 *Drosophila* genome sequences will serve as a powerful tool for gleaning further insight into genetic, developmental, regulatory and evolutionary processes.

Table 2 | A summary of annotated features across all 12 genomes

	Protein-coding gene annotations			Non-coding RNA annotations				Repeat coverage (%)*	Genome size (Mb; assembly†/flow cytometry‡)
	Total no. of protein- coding genes (per cent with *D. melanogaster* homologue)	Coding sequence/ intron (Mb)	tRNA (pseudo)	snoRNA	miRNA	rRNA (5.8S + 5S)	snRNA		
D. melanogaster	13,733 (100%)	38.9/21.8	297 (4)	250	78	101	28	5.35	118/200
D. simulans	15,983 (80.0%)	45.8/19.6	268 (2)	246	70	72	32	2.73	111/162
D. sechellia	16,884 (81.2%)	47.9/21.9	312 (13)	242	78	133	30	3.67	115/171
D. yakuba	16,423 (82.5%)	50.8/22.9	380 (52)	255	80	55	37	12.04	127/190
D. erecta	15,324 (86.4%)	49.1/22.0	286 (2)	252	81	101	38	6.97	134/135
D. ananassae	15,276 (83.0%)	57.3/22.3	472 (165)	194	76	134	29	24.93	176/217
D. pseudoobscura	16,363 (78.2%)	49.7/24.0	295 (1)	203	73	55	31	2.76	127/193
D. persimilis	17,325 (72.6%)	54.0/21.9	306 (1)	199	75	80	31	8.47	138/193
D. willistoni	15,816 (78.8%)	65.4/23.5	484 (164)	216	77	76	37	15.57	187/222
D. virilis	14,680 (82.7%)	57.9/21.7	279 (2)	165	74	294	31	13.96	172/364
D. mojavensis	14,849 (80.8%)	57.8/21.9	267 (3)	139	71	74	30	8.92	161/130
D. grimshawi	15,270 (81.3%)	54.9/22.5	261 (1)	154	82	70	32	2.84	138/231

* Repeat coverage calculated as the fraction of scaffolds >200 kb covered by repeats, estimated as the midpoint between BLASTER-tx + PILER and RepeatMasker + ReAS (Supplementary Information section 3). †Total genome size estimated as the sum of base pairs in genomic scaffold >200,000 bp. ‡Genome size estimates based on flow cytometry[38].

Ancestral genome reconstructions

Million years	Karyotype 	Karyotype + Gene order 	Genomic Sequence TCGAGGACTTAA GCATGTC
530	**Chordate** (Putnam et al. 2008)		
470	**Vertebrate** (Kohn et al. 2006) (Nakatani et al. 2007)		
310	**Amniota** (Nakatani et al. 2007)		
250	**Teleostei** (Jaillon et al. 2004) (Kasahara et al. 2007)		
90	**(Boreo)Eutheria** (Ruiz-Herrera 2012)	**Boreoeutheria** (Bourque et al. 2004) (Ma et al. 2006) (Chauve & Tannier 2008)	**Boreoeutheria** (Blanchette et al. 2004) (Paten et al. 2008)

Decoding Plant and Animal Genome Plasticity from Differential Paleo-Evolutionary Patterns and Processes

Florent Murat[1], Yves Van de Peer[2,3], and Jérôme Salse[1,*]

Continuing advances in genome sequencing technologies and computational methods for comparative genomics currently allow inferring the evolutionary history of entire plant and animal genomes. Based on the comparison of the plant and animal genome paleohistory, major differences are unveiled in 1) evolutionary mechanisms (i.e., polyploidization versus diploidization processes), 2) genome conservation (i.e., coding versus noncoding sequence maintenance), and 3) modern genome architecture (i.e., genome organization including repeats expansion versus contraction phenomena). This article discusses how extant animal and plant genomes are the result of inherently different rates and modes of genome evolution resulting in relatively stable animal and much more dynamic and plastic plant genomes.

Table 2

Major Differences in Plant and Animal Genome Structure, Function, and Evolution

Genome Properties	Features	Plants (monocots and dicots)	Animals (vertebrates)	References
Ancestor				
	Protochromosomes	5–7	10–12	CA
	Protogenes	~10,000–15,000	~13,000–20,000	CA
	Gene space size	~25 Mb	~50 Mb	CA
Structure				
	Chromosome/genomes	Shuffled	Stable	CA
	Genes (size, exon size, exon number)	2.9 Kb/384 bp/4.7	39.7 Kb/290 bp/8.5	CA
	CNS	Short/less conserved	Long/highly conserved	Reineke et al. (2011)
	Gene families	Less/genome wide	Numerous/tandem	Kejnovsky et al. (2009)
	TE	Mainly class I LTR / recent	Mainly class I non-LTR / old	CA
Function				
	Neo/sub-functionalization	High between duplicates	Low	Pont et al. (2011)
	Splice variant	Low	High	Taher et al. (2011)
	Small RNA	miRNA/target coevolution	miRNA emergence/new target	Axtell et al. (2011)
Evolution				
	Duplication/polyploidy	Frequent/recent	Rare/old	CA
	Fusion	Centromeric-based	Telomeric-based	CA
	Recombination	High/variable	Low/stable	Gaut et al. (2007)
Plants versus animals	Chromosomes and genomes	Plastic	Stable	CA

NOTE.—CA, current analysis; compared with those discussed from the literature (references cited).

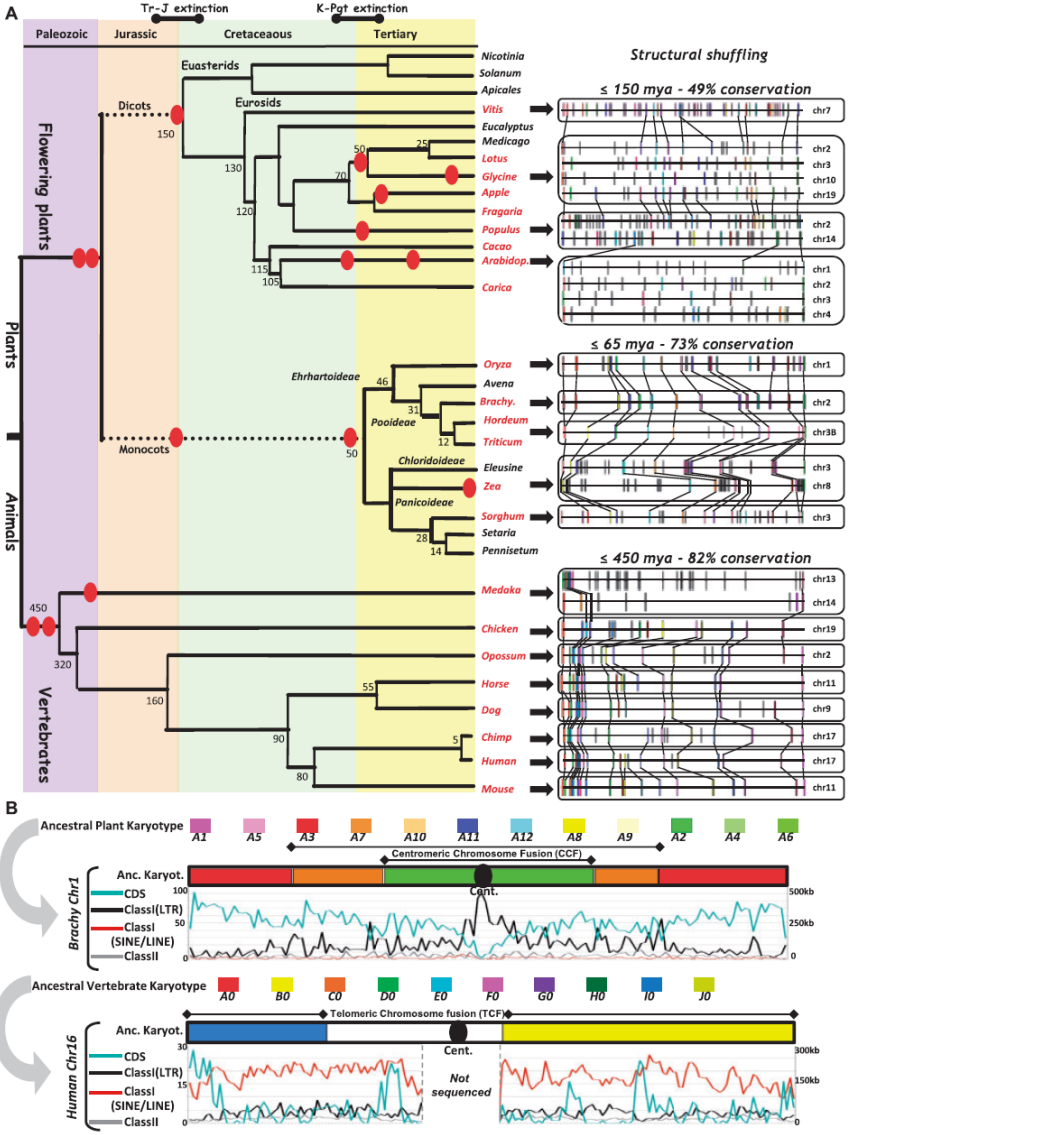


FIG. 2.—Gene conservation in plant and animal genomes. (*A*) A highly pruned phylogenetic tree of the plants and animals is shown at the left end side of the figure with speciation dates of the branches (in millions years) and duplication events highlighted as red dots. Micro-synteny conservation is shown at the right end side of the figure with homologous genes in the same color code and connected with black lines. (*B*) Comparison of plant (*Brachypodium* chromosome 1) and animal (human chromosome 16) genome heat maps. Each chromosome structure is illustrated based on the ancestral karyotype (10 and 12 colour codes, respectively for animals and plants) and associated with corresponding CDS (blue), TE class I LTR (black), TE class I non-LTR (purple) and TE class II (gray) distribution curves. Within 500-kb-sized windows covering the entire considered chromosome, CDS distribution (left) represents the number of annotated genes and TEs distribution (right, Y-axis) represents the cumulative size in "Kb" covered by either class I (black curve) and class II (gray curve) elements.

The others: our biased perspective of eukaryotic genomes

Javier del Campo[1,2], Michael E. Sieracki[3], Robert Molestina[4], Patrick Keeling[2], Ramon Massana[5], and Iñaki Ruiz-Trillo[1,6,7]

[1] Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[2] University of British Columbia, Vancouver, BC, Canada
[3] Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA
[4] American Type Culture Collection, Manassas, VA, USA
[5] Institut de Ciències del Mar, CSIC, Barcelona, Catalonia, Spain
[6] Departament de Genètica, Universitat de Barcelona, Barcelona, Catalonia, Spain
[7] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

Understanding the origin and evolution of the eukaryotic cell and the full diversity of eukaryotes is relevant to many biological disciplines. However, our current understanding of eukaryotic genomes is extremely biased, leading to a skewed view of eukaryotic biology. We argue that a phylogeny-driven initiative to cover the full eukaryotic diversity is needed to overcome this bias. We encourage the community: (i) to sequence a representative of the neglected groups available at public culture collections, (ii) to increase our culturing efforts, and (iii) to embrace single cell genomics to access organisms refractory to propagation in culture. We hope that the community will welcome this proposal, explore the approaches suggested, and join efforts to sequence the full diversity of eukaryotes.

Eukaryotes are the most complex of the three domains of life. The origin of eukaryotic cells and their complexity remains one of the longest-debated questions in biology, famously referred to by Roger Stanier as the 'greatest single evolutionary discontinuity' in life [1]. Thus, understanding how this complex cell originated and how it evolved into the diversity of forms we see today is relevant to all biological disciplines including cell biology, evolutionary biology, ecology, genetics, and biomedical research. Progress in this area relies heavily on both genome data from extant organisms and on an understanding of their phylogenetic relationships.

Genome sequencing is a powerful tool that helps us to understand the complexity of eukaryotes and their evolutionary history. However, there is a significant bias in eukaryotic genomics that impoverishes our understanding of the diversity of eukaryotes, and leads to skewed views of

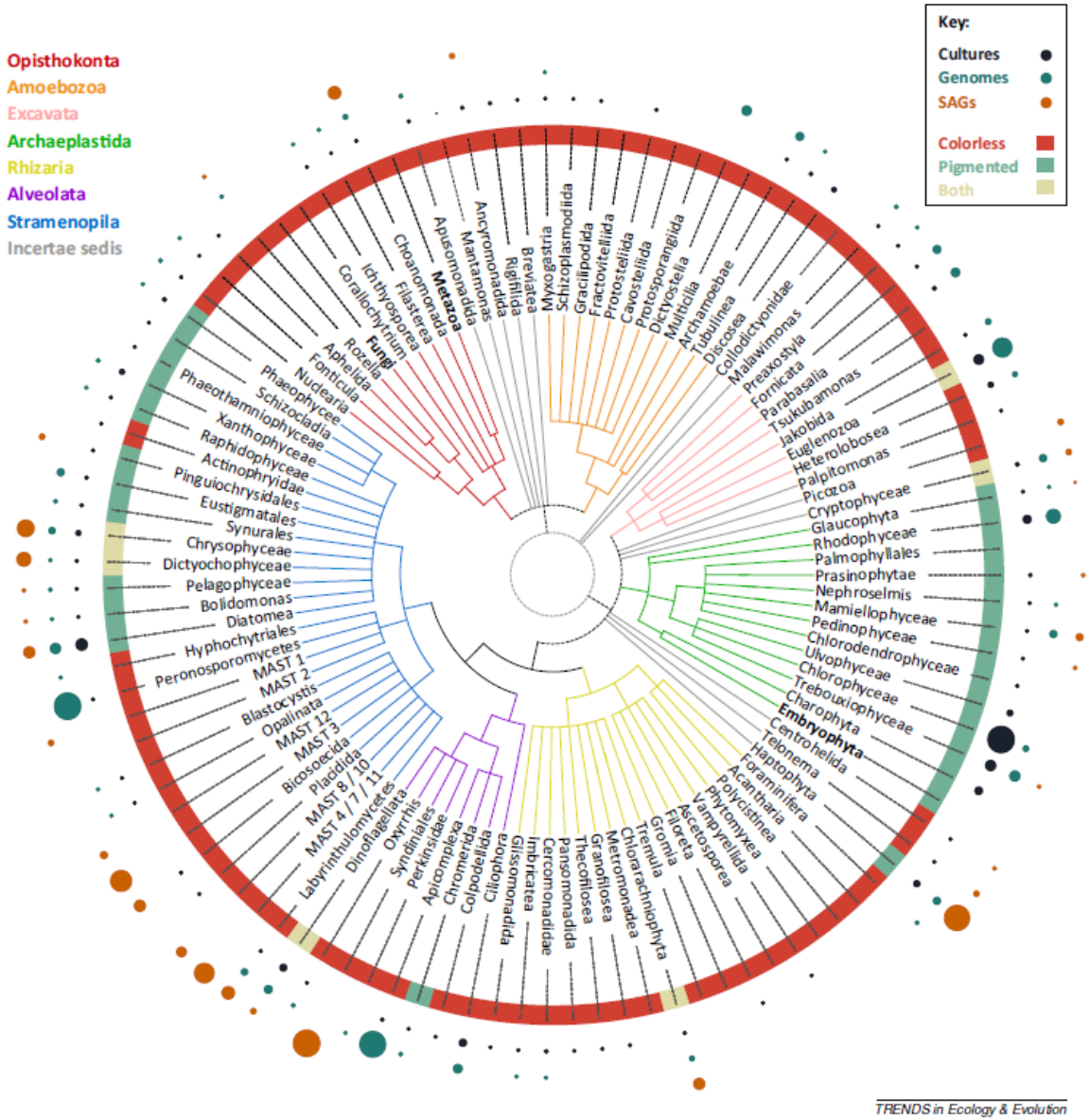


Figure 4. The tree of eukaryotes, showing the distribution of current effort on culturing, genomics, and environmental single amplified genome (SAG) genomics for the main protistan lineages. Eukaryotic schematic tree representing major lineages. Colored branches represent the seven main eukaryotic supergroups, whereas grey branches are phylogenetically contentious taxa. The sizes of the dots indicate the proportion of species/OTU_{97} in each database. Culture data are from the analyzed publicly available protist culture collections ($n = 3084$). Genome data were extracted from the Genomes OnLine Database (GOLD) ($n = 258$) [9]. SAGs of OTU_{97} correspond to those retrieved during the Tara Oceans cruise ($n = 158$) (M.E.S., unpublished data). Taxonomic annotation of all datasets is based on [28]. The 'big three' (in bold) have been excluded from this analysis. Abbreviation: OTU_{97}, operational taxonomic unit (>97% sequence identity).