Genomska pokrajina (=struktura genoma) in evolucija genomske komplesnosti

Genomic landscape = Genome Structure

Genome Structure (human)

This section describes several of the ***noncoding attributes of the assembled genome sequence and their correlations with the predicted gene set***. These include an ***analysis of GC content*** and ***gene density*** in the context of cytogenetic maps of the genome, an ***enumerative analysis of CpG islands***, and a brief description of the ***genome-wide repetitive elements***.

Genome landscape

First, we analyse the phylogenetic position of **platypus** and confirm that marsupials and eutherians are more closely related than either is to monotremes. We then describe platypus chromosomes and observe some properties of platypus ***interspersed and tandem repeats.*** We also discuss a ***potential relationship between interspersed repeats and genomic imprinting*** and investigate how the extremely high ***GC fraction*** in platypus affects the strong association seen in eutherians between ***CpG islands and gene promoters***.

Long-range variation in GC content

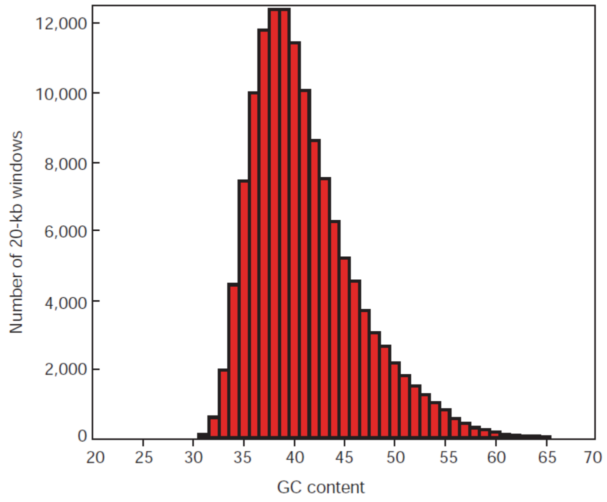
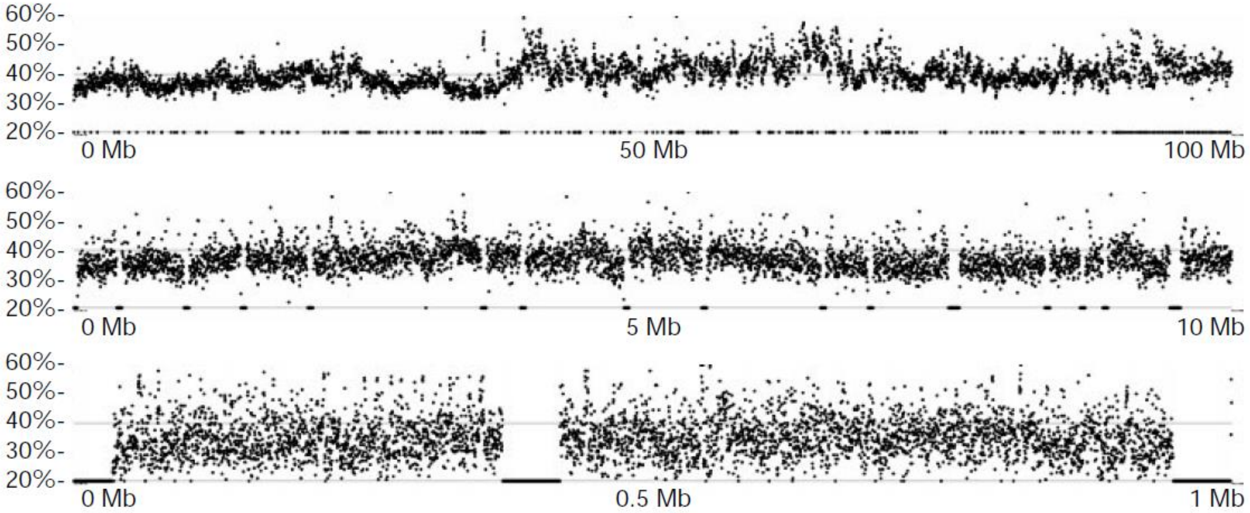


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.



CpG islands

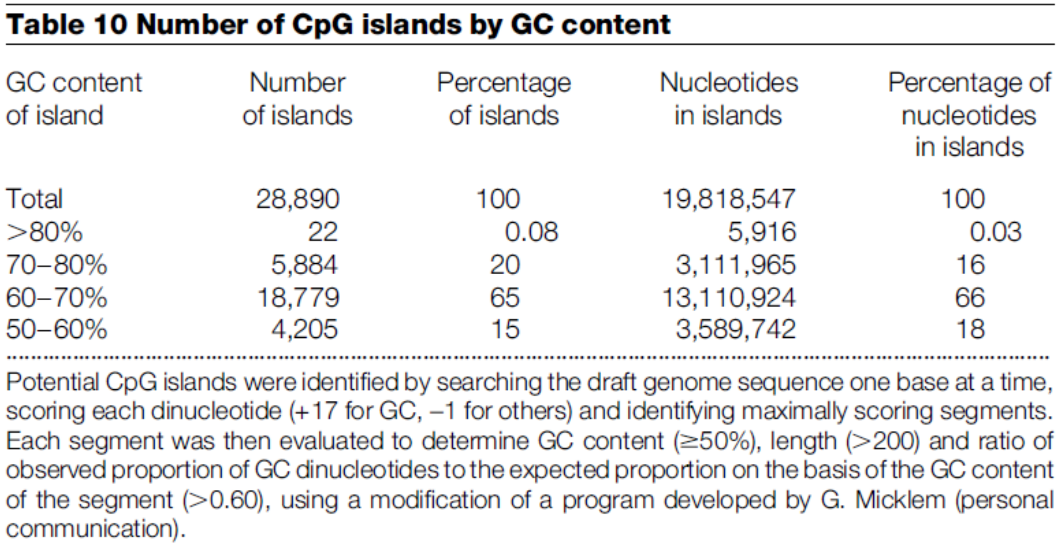


Table 10 Number of CpG islands by GC content

GC content of island	Number of islands	Percentage of islands	Nucleotides in islands	Percentage of nucleotides in islands
Total	28,890	100	19,818,547	100
>80%	22	0.08	5,916	0.03
70–80%	5,884	20	3,111,965	16
60–70%	18,779	65	13,110,924	66
50–60%	4,205	15	3,589,742	18

Potential CpG islands were identified by searching the draft genome sequence one base at a time, scoring each dinucleotide (+17 for GC, −1 for others) and identifying maximally scoring segments. Each segment was then evaluated to determine GC content (≥50%), length (>200) and ratio of observed proportion of GC dinucleotides to the expected proportion on the basis of the GC content of the segment (>0.60), using a modification of a program developed by G. Micklem (personal communication).

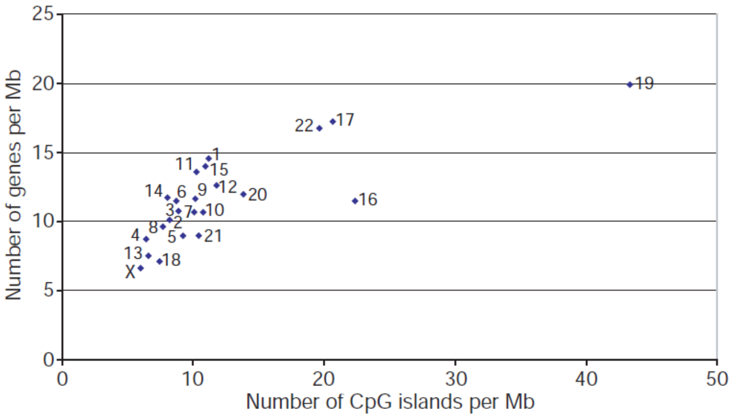


Figure 14 Number of CpG islands per Mb for each chromosome, plotted against the number of genes per Mb (the number of genes was taken from GeneMap98 (ref. 100)). Chromosomes 16, 17, 22 and particularly 19 are clear outliers, with a density of CpG islands that is even greater than would be expected from the high gene counts for these four chromosomes.

Isochores, GC content and gene density

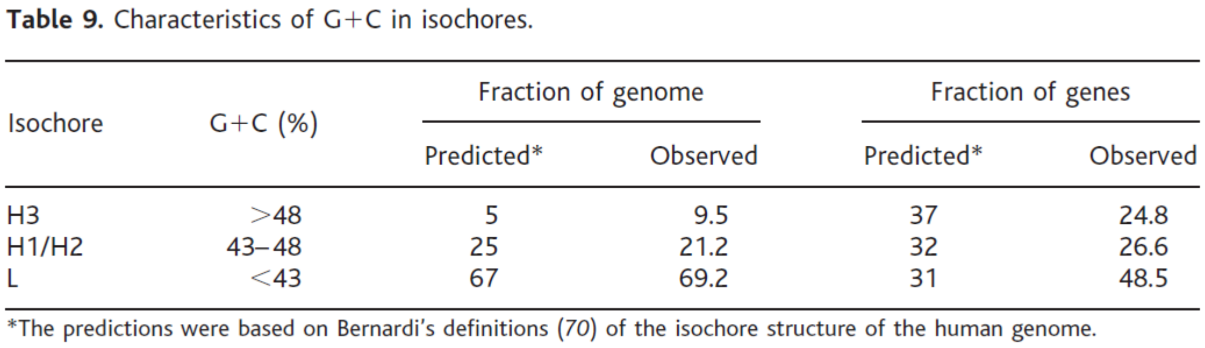
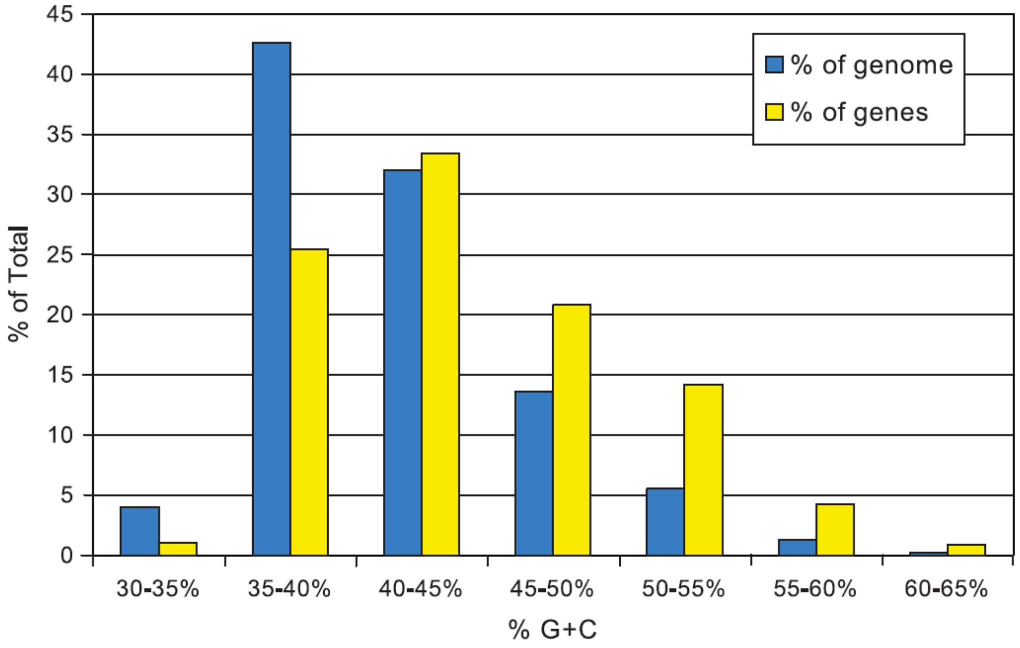


Table 9. Characteristics of G+C in isochores.

Isochore	G+C (%)	Fraction of genome		Fraction of genes	
		Predicted*	Observed	Predicted*	Observed
H3	>48	5	9.5	37	24.8
H1/H2	43–48	25	21.2	32	26.6
L	<43	67	69.2	31	48.5

*The predictions were based on Bernardi's definitions (*70*) of the isochore structure of the human genome.



Relation between GC content and gene density. The blue bars show the percent of the genome (in 50-kbp windows) with the indicated GC content. The percent of the total number of genes associated with each GC bin is represented by the yellow bars. The graph shows that about 5% of the genome has a GC content of between 50 and 55%, but that this portion contains nearly 15% of the genes.

Repeat content of the human genome

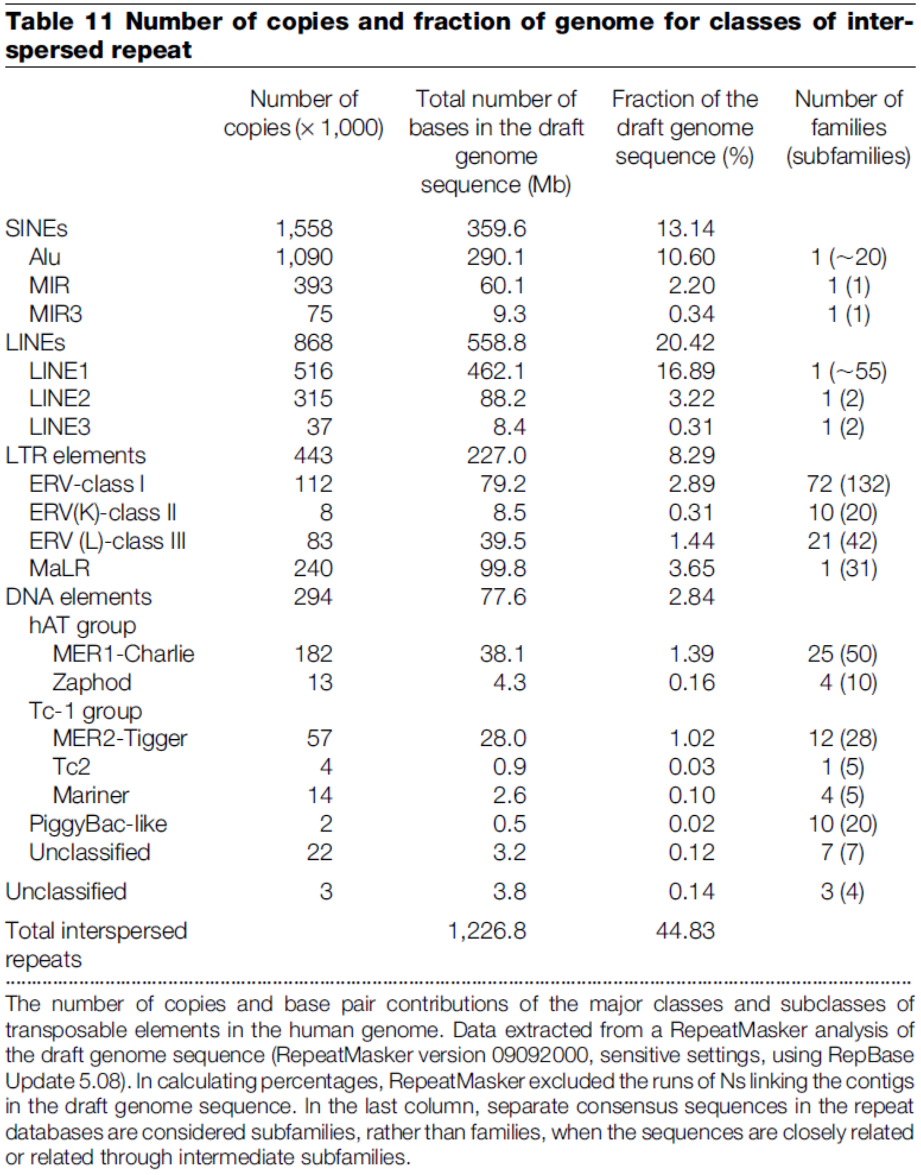
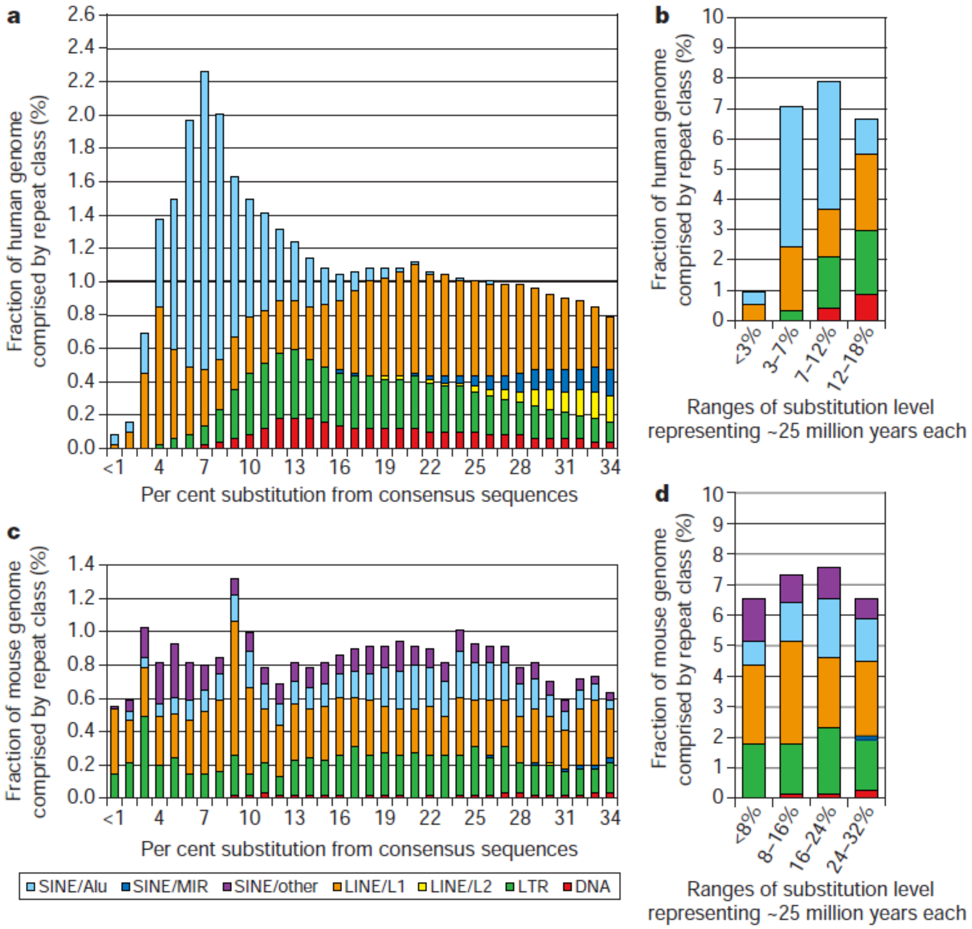


Table 11 Number of copies and fraction of genome for classes of interspersed repeat

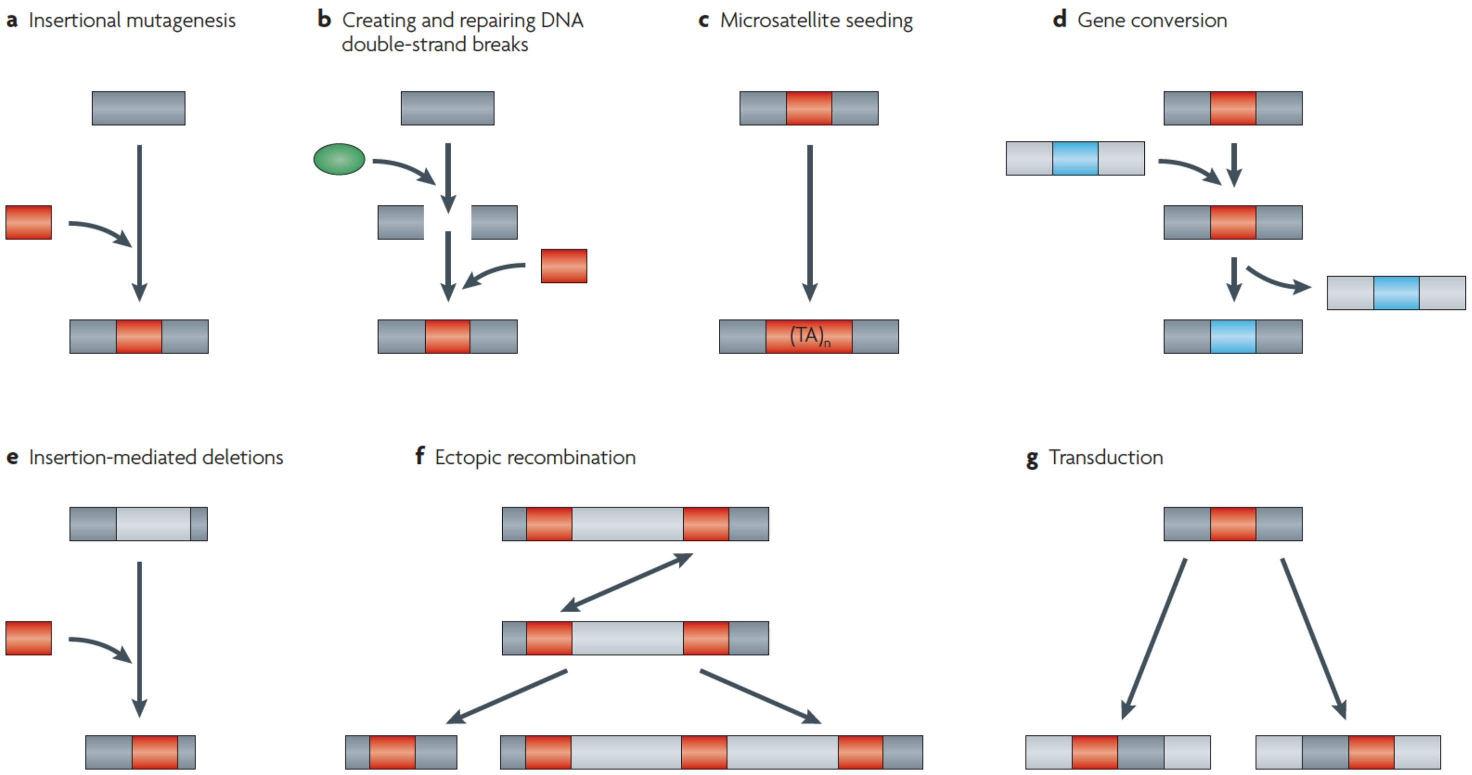
	Number of copies (× 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.



Age distribution of interspersed repeats in the human and mouse genomes. Bases covered by interspersed repeats were sorted by their divergence from their consensus sequence (which approximates the repeat's original sequence at the time of insertion).

Impact of retrotransposons on genome structure


a Insertional mutagenesis
b Creating and repairing DNA double-strand breaks
c Microsatellite seeding
(TA)n
d Gene conversion
e Insertion-mediated deletions
f Ectopic recombination
g Transduction

Impact of retrotransposons on human genome structure.

a | Typical insertion of a LINE-1 (L1), Alu or SVA retrotransposon (red box) at a new genomic site (dark grey). If the new genomic site is a genic region, the retrotransposon may cause insertional mutagenesis.

b | The protein product (green oval) of an L1 element may create DNA double-strand breaks (broken dark grey area). Alternatively, an existing double-strand break may be repaired by non-classical endonuclease-independent insertion of a retrotransposon.

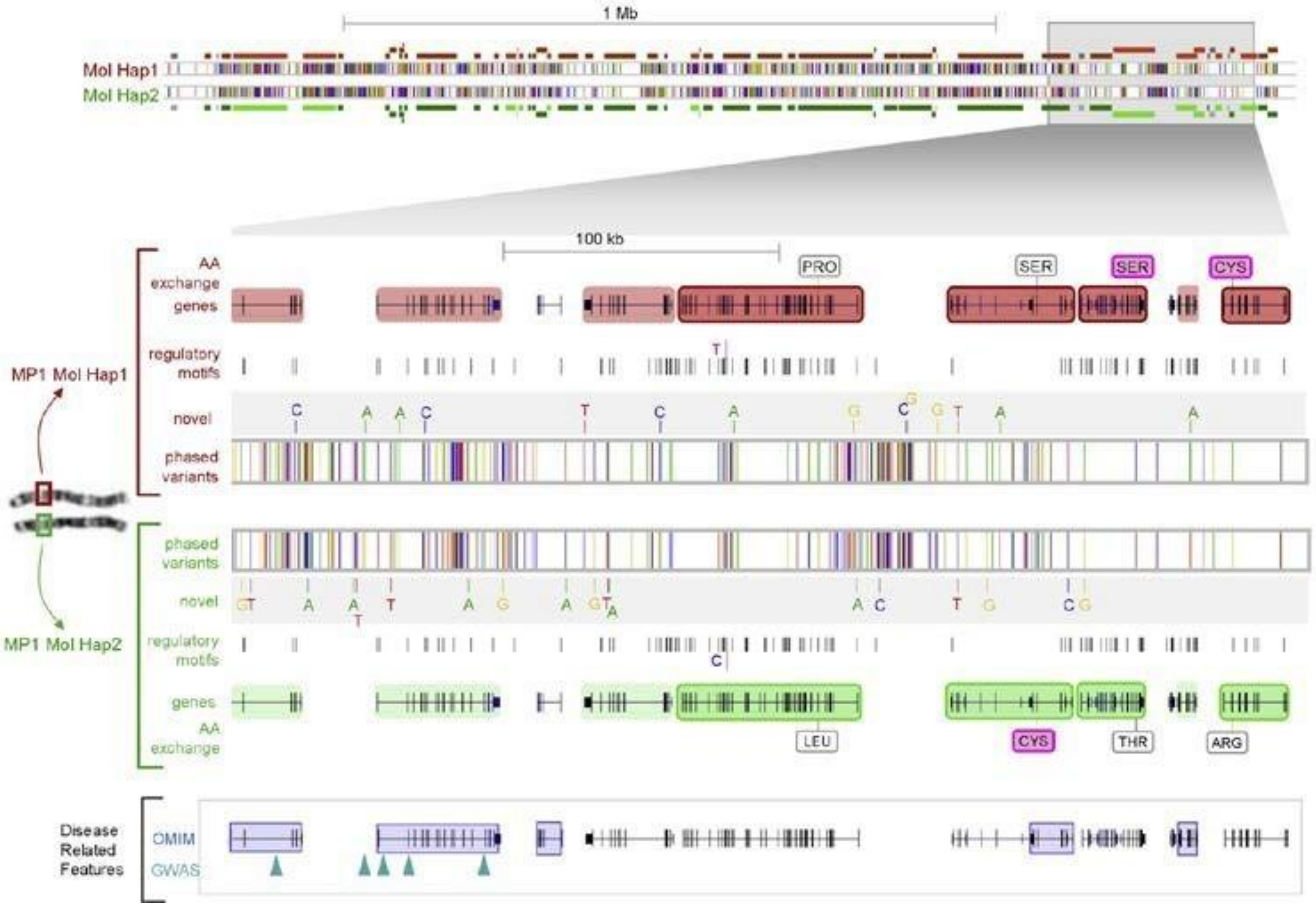
c | Microsatellites (for example, (TA)n) may arise from the homopolymeric tracts that are endogenous to retrotransposons.

d | Gene conversion may alter the sequence compositions of homologous retrotransposon copies (red and blue boxes).

e | The insertion of a retrotransposon is sometimes associated with the concomitant deletion of a target genomic sequence (light grey box).

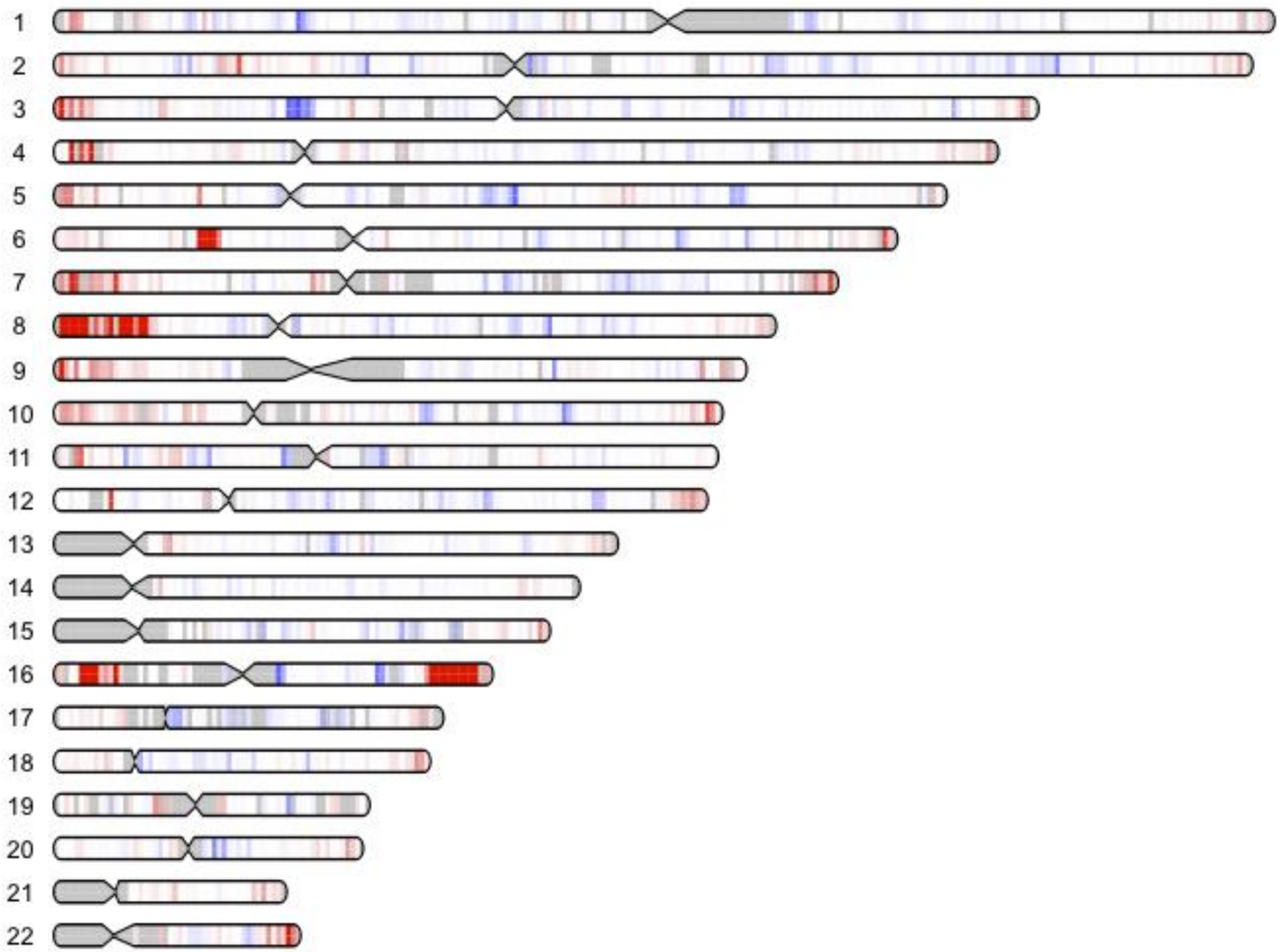
f | Ectopic recombination (double arrowhead) between non-allelic homologous retrotransposons may result in genomic rearrangements, such as deletions (left) or duplications (right) of intervening genomic sequences.

g | During the duplication of a retrotransposon, the downstream 3' flanking sequence or the upstream 5' flanking sequence (dark grey boxes) may also be duplicated (known as 3' or 5' transduction, respectively). This results in the retrotransposition of the 3' flanking sequence (left) or the 5' flanking sequence (right) along with the retrotransposon.

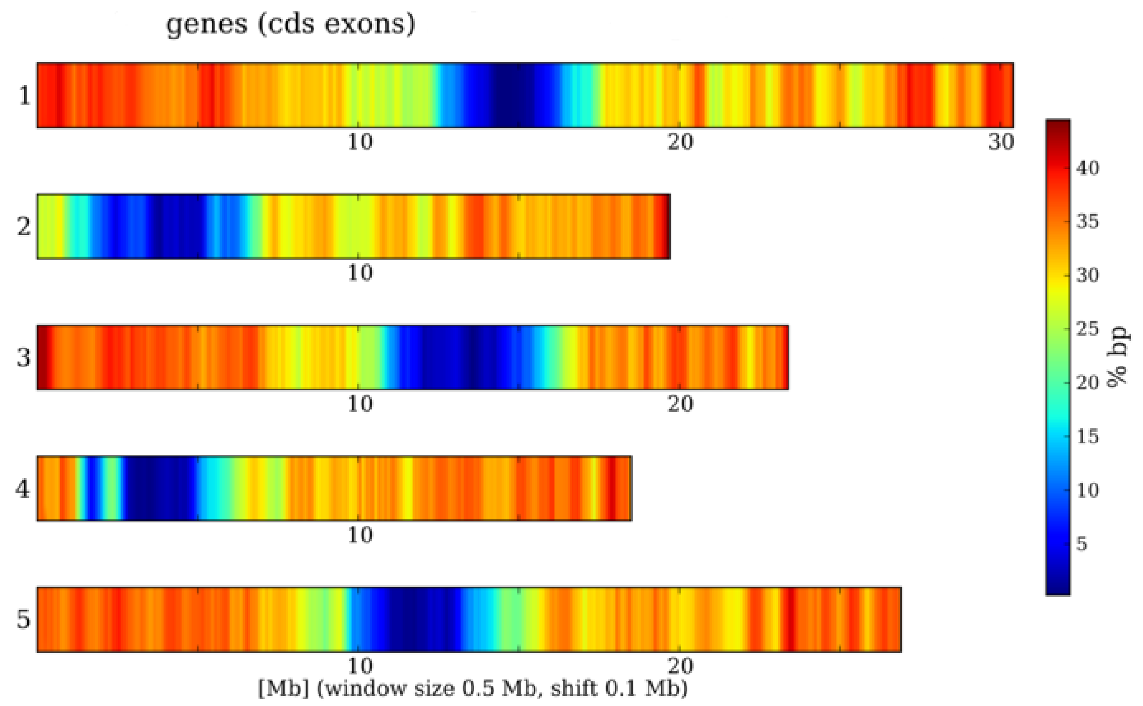


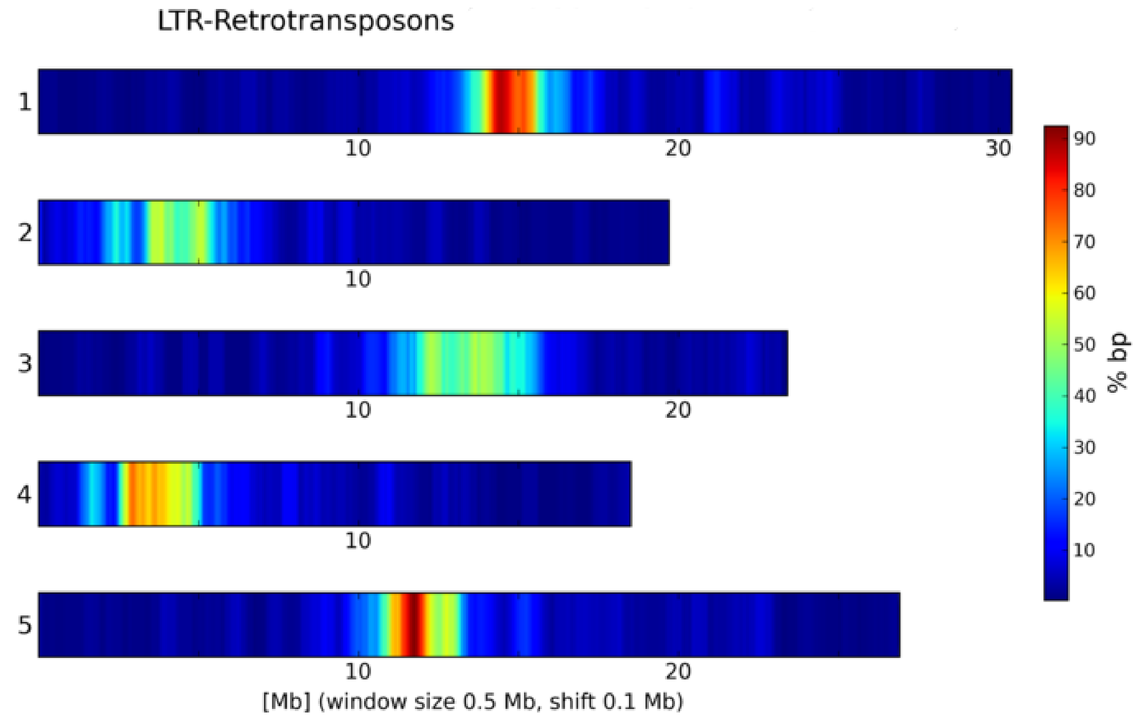
Determination of haploid landscapes of functional variation - towards phase-sensitive personal genomics.

The determination of contiguous molecular haplotype sequences in the Megabase range (‘haploid landscapes’) is crucial to translate individual genomic variation into the functionally active proteome. These may encode functional entities that affect gene expression in a coordinated way. ***Example of a Megabase-size haploid landscape of functional variation on chromosome 19***. Differences between the two molecular haplotypes are shown at the ***nucleotide level, regulatory level, and level of genome organization*** (from Suk et al., Genome Res 2011; 21(10):1672-85).



Genomic distribution of SNP density on human chromosomes 1 to 22. The colours show the SNP density, with **red** indicating higher densities and **blue** indicating lower densities. Note **high rates of SNP variation near the ends of the chromosomes**.

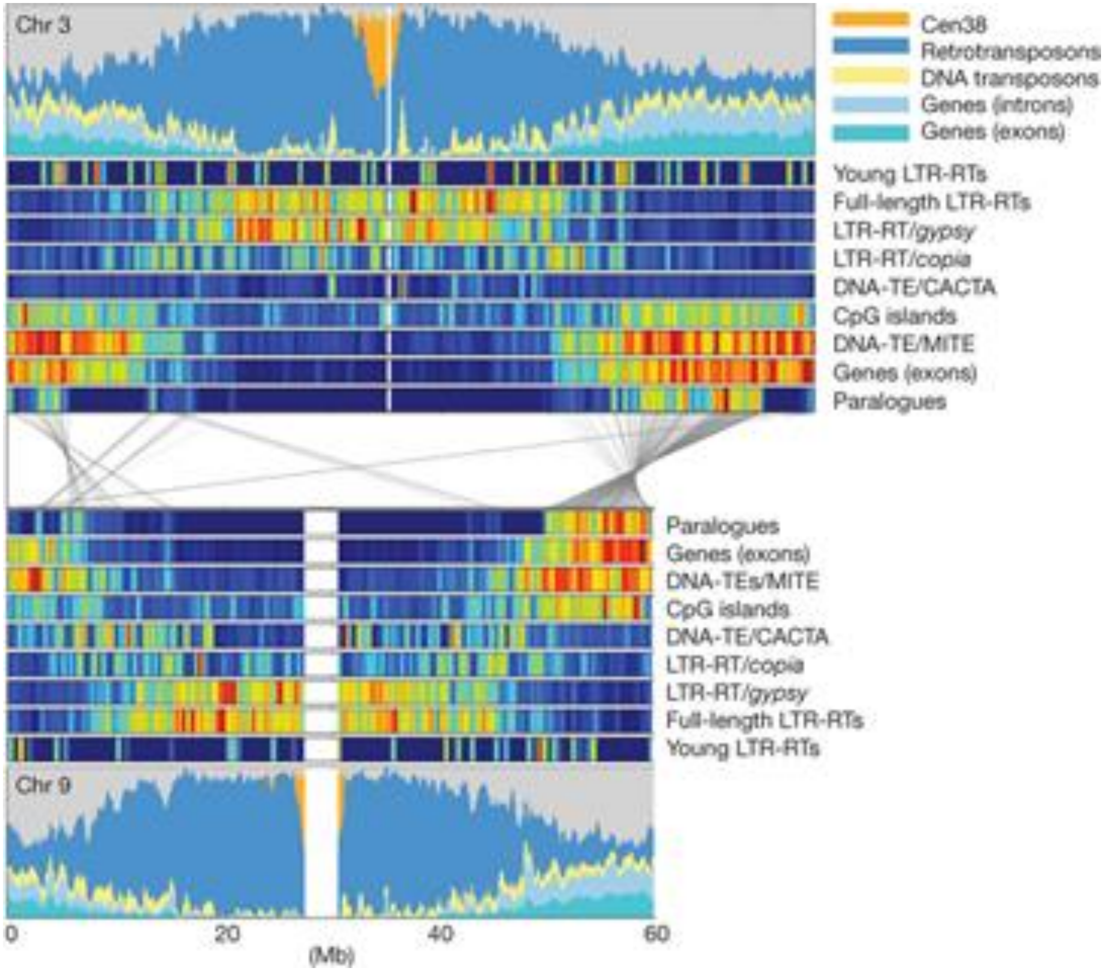




Chromosomal Heatmaps. ***For the graphical representation of <u>position and density of sequence features</u> we use heatmaps***. They graphically represent density of the respective feature by colors. Here, the color-coded features (**repeats, genes**) are shown for Arabidopsis thaliana chromosomes. **Blue** represents lowest (0%), **green** medium (50%) and **red** highest (100%) density of the respective feature. Relative density has been determined within a sliding window of 0.5 MB size and in steps of 0.1 MB.

A **heat map** is a graphical representation of data where the individual values contained in a matrix are represented as colors.//The easiest way to understand a heat map is to think of a table or spreadsheet which contains colors instead of numbers.

Heat map = toplotni graf; „toplotna slika“



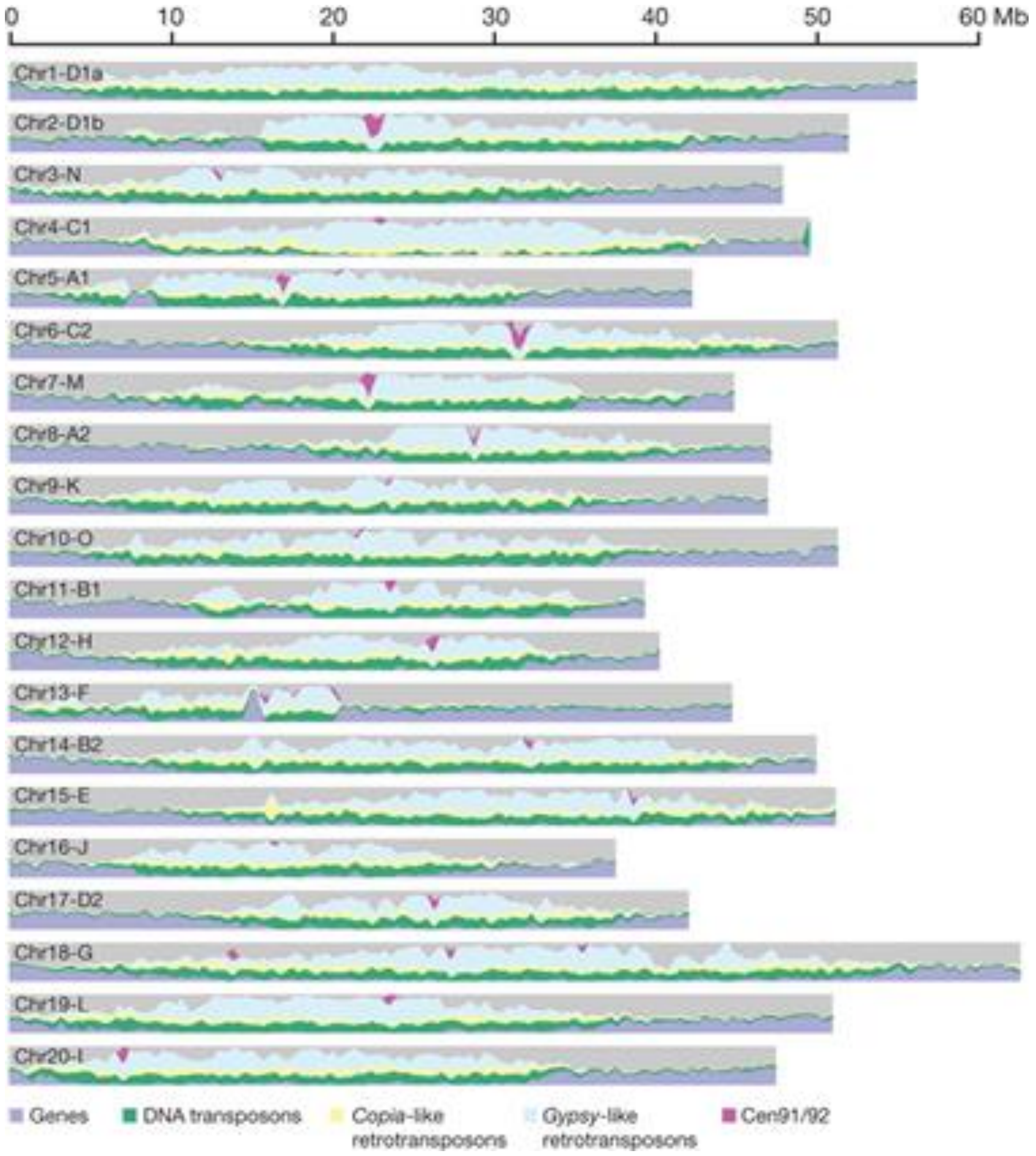
Genomic landscape of sorghum chromosomes 3 and 9.

Area charts quantify ***retrotransposons (55%), genes (6% exons, 8% introns), DNA transposons (7%) and centromeric repeats (2%).***

Lines between chromosomes 3 and 9 connect ***collinear duplicated genes***.

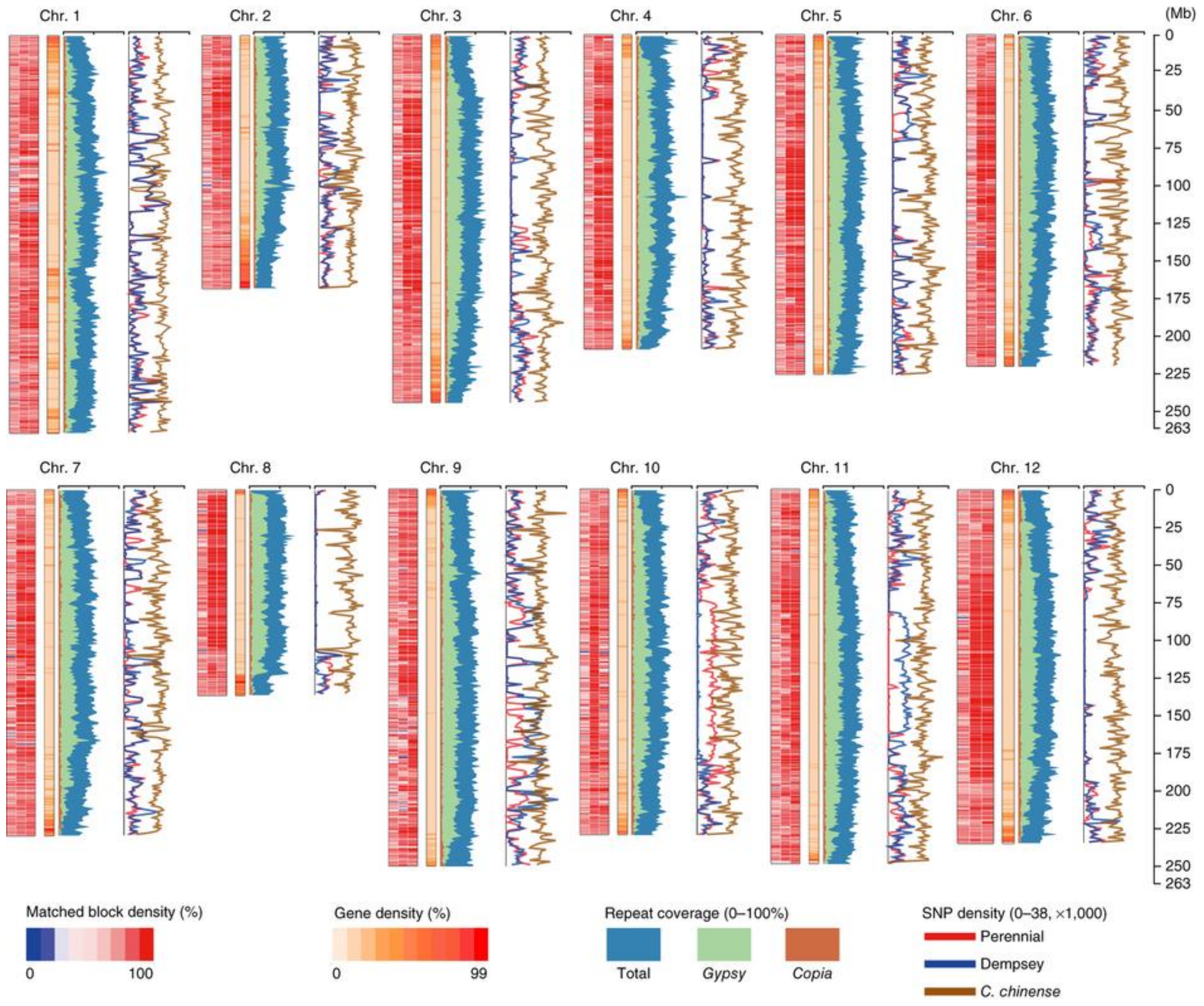
Heat-map tracks detail the distribution of selected elements.

Cen38, sorghum-specific ***centromeric*** repeat10;
RTs, retrotransposons (class I);
LTR-RTs, long terminal repeat retrotransposons;
DNA-TEs, DNA transposons (class II).



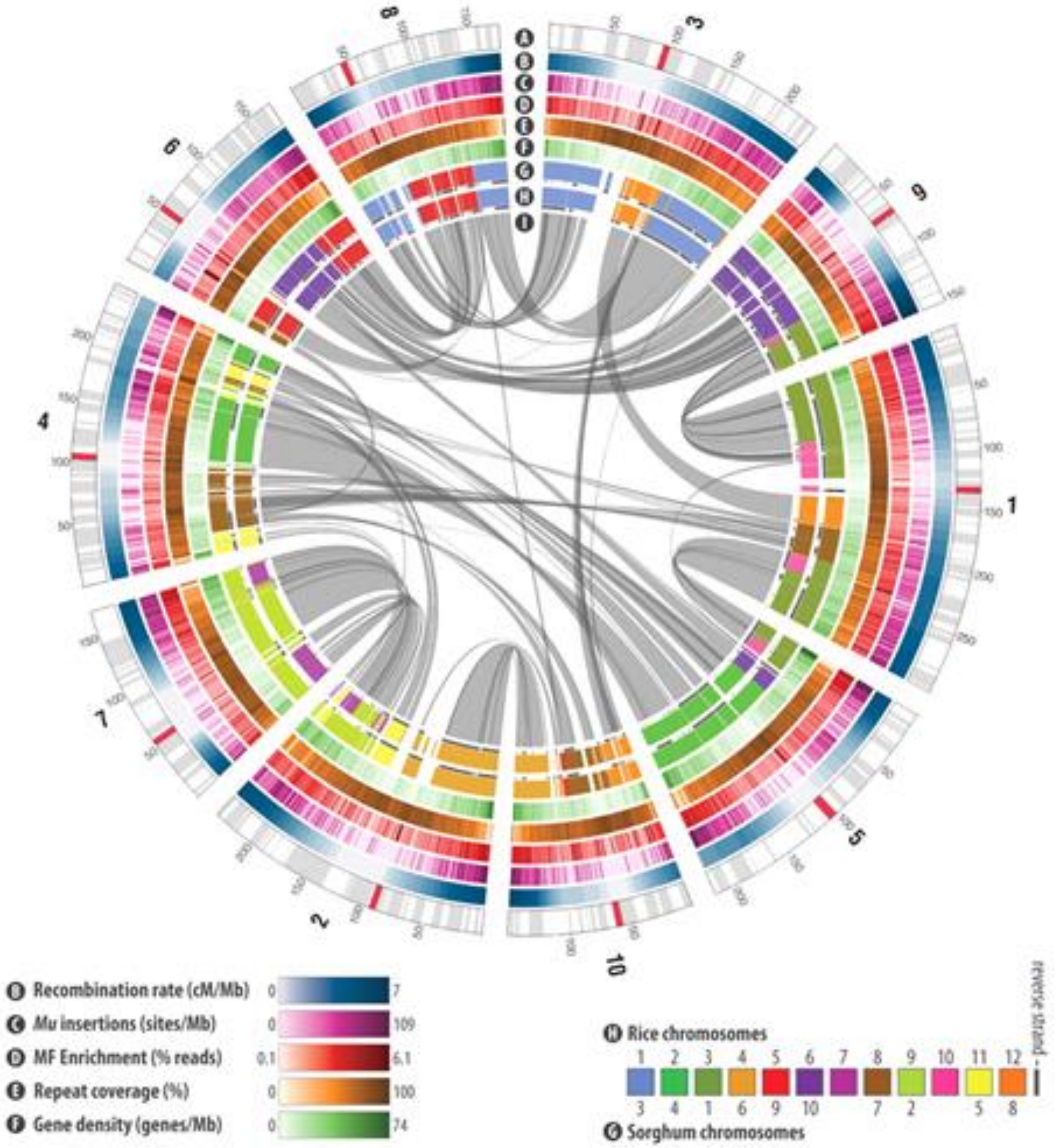
Genomic landscape of the 20 assembled soybean chromosomes.

Major DNA components are categorized into *genes (blue),* *DNA transposons (green),* *Copia-like retrotransposons (yellow), Gypsy-like retrotransposons (cyan) and Cent91/92 (a soybean-specific centromeric repeat (pink))*, with respective DNA contents of *18%, 17%, 13%, 30% and 1% of the genome sequence*. Unclassified DNA content is coloured grey. Categories were determined for 0.5-Mb windows with a 0.1-Mb shift.



Genomic landscape of pepper chromosomes. Left to right: density of matched blocks, gene density, repeat coverage and SNP density. Density of matched blocks is presented for C. chinense, Dempsey and Perennial (left to right) for 500-kb windows. ***Gene density is presented as the number of genes within 1-Mb intervals.*** Coverage by repeats represents the proportion of total TEs, Gypsy elements and Copia elements of LTRs within 1-Mb intervals. SNP density is presented as the number of SNPs per 1-Mb interval.

The B73 Maize Genome: Complexity, Diversity, and Dynamics.


A
B
C
D
E
F
G
H
I
1
2
3
4
5
6
7
8
9
10
B Recombination rate (cM/Mb) 0 7
C Mu insertions (sites/Mb) 0 109
D MF Enrichment (% reads) 0.1 6.1
E Repeat coverage (%) 0 100
F Gene density (genes/Mb) 0 74
H Rice chromosomes
1 2 3 4 5 6 7 8 9 10 11 12
3 4 1 6 9 10 7 2 5 8
G Sorghum chromosomes
reverse strand

The maize B73 reference genome (B73 RefGen_v1): Concentric circles show aspects of the genome.

Chromosome structure (A). Reference chromosomes with physical fingerprint contigs as alternating gray and white bands. Presumed centromeric positions are indicated by red bands; enlarged for emphasis.

Genetic map (B). Genetic linkage across the genome, on the basis of 6363 genetically and physically mapped markers.

Mu insertions (C). Genome mappings of nonredundant Mu insertion sites.

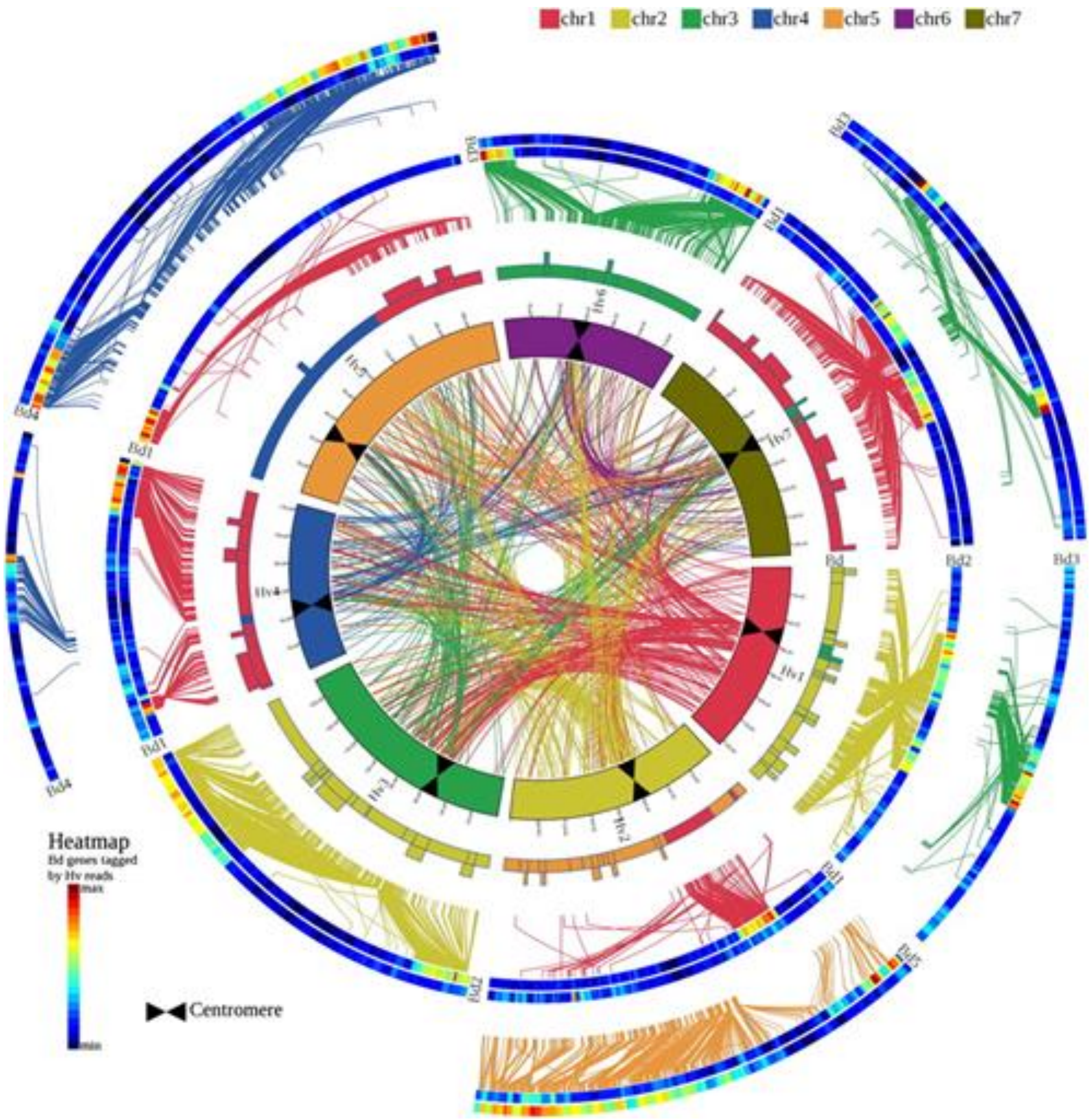
Methyl-filtration reads (D). Enrichment and depletion of methyl filtration. For each nonoverlapping 1-Mb window, read counts were divided by the total number of mapped reads.

Repeats (E). Sequence coverage of TEs with RepeatMasker with all identified intact elements in maize.

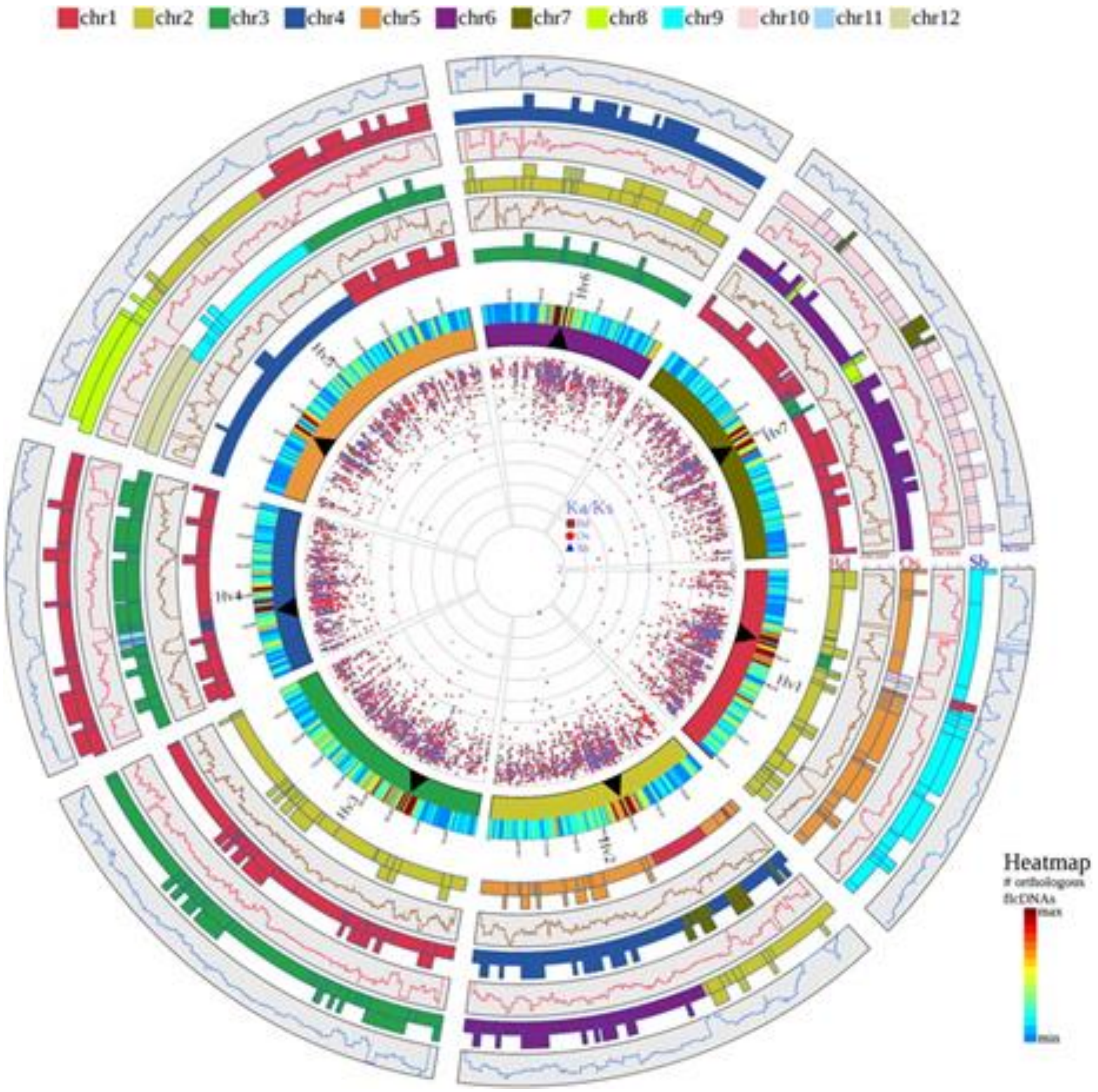
Genes (F). ***Density of genes*** in the filtered gene set across the genome, from a ***gene count per 1-Mb sliding window at 200-kb intervals***.

Sorghum synteny (G) and **rice synteny (H).** Syntenic blocks between maize and related cereals on the basis of 27,550 gene orthologs. Underlined blocks indicate alignment in the reverse strand.

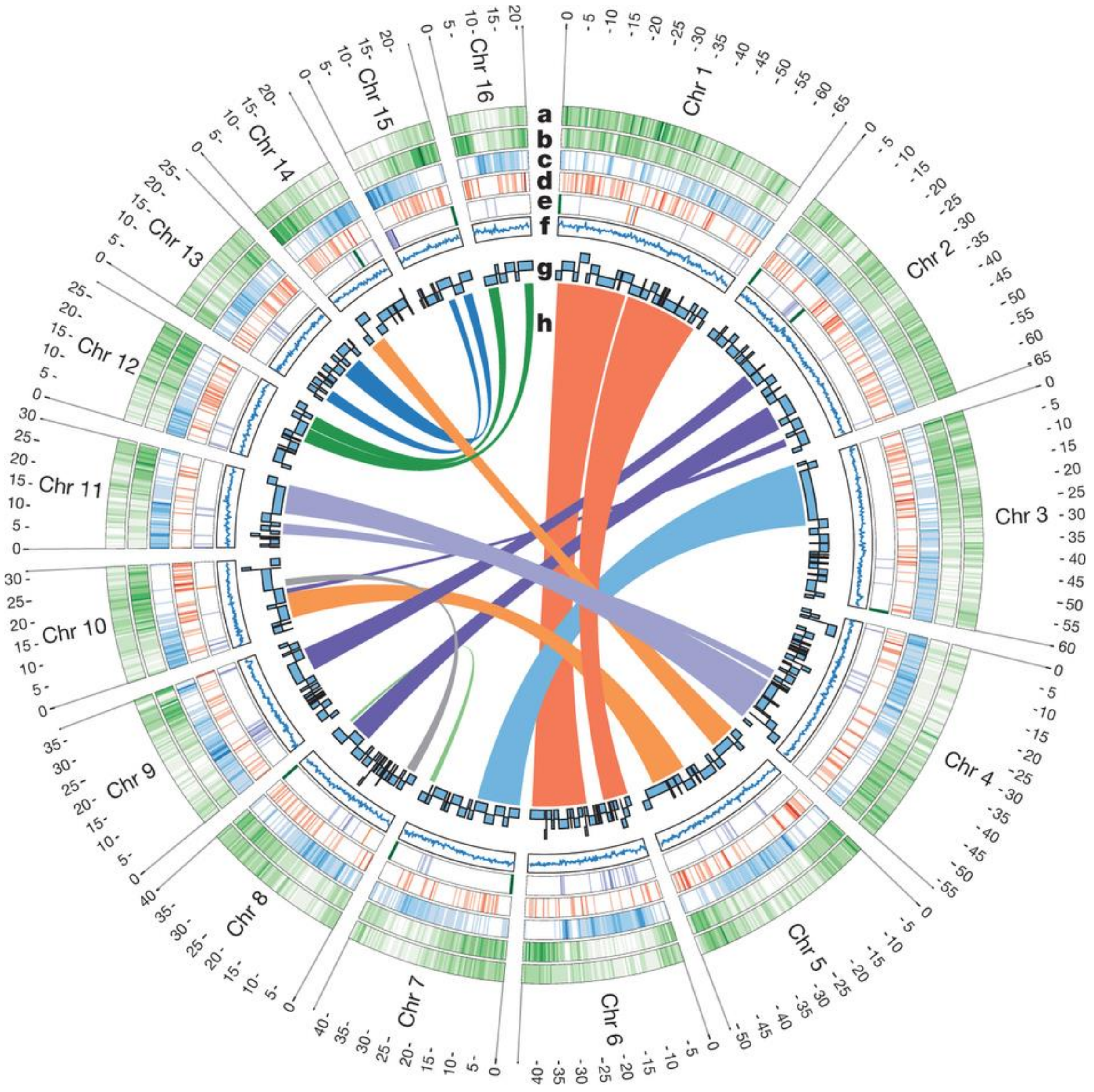
Homoeology map (I). Oriented homoeologous sites of duplicated gene blocks within maize.



High-Resolution Comparative Analysis between Barley and B. distachyon. ***High-density comparative analysis of the linear gene order*** of the barley genome zippers versus the sequenced model grass genome of Brachypodium. The figure includes **four sets of concentric circles**: the ***inner circle*** represents the seven chromosomes of barley scaled according to the barley genetic map (bars at 10-cM intervals). Each barley chromosome is assigned a color according to the sequence on the color key, starting with chr1 through chr7. The positions of the barley centromeres are indicated by black bars. Moving outwards, the **second circle** illustrates a schematic model of the seven barley chromosomes, but this time color-coded according to ***blocks of conserved synteny with the model genome***. The color coding is again based on the sequence on the color key, but this time is based on the model genome linkage groups, starting with chr1 through chr5 for Brachypodium. *Boxes extending from these colored bars indicate regions involved in larger-scale structural changes (e.g., inversions)*. The *outer partially complete circles of heat map* colored bars represent pseudomolecules of the model genome linkage groups arranged according to ***conserved synteny with barley 1H-7H***. When pairs of adjacent heat map bars are shown, they illustrate where the homologs of a short (inner heat map bar) or a long (outer heat map bar) barley chromosome arm data set is allocated to the respective model genome pseudochromosome. The **heat maps illustrate the density of genes** hit by the 454 shotgun reads from the relevant barley chromosome arm. ***Conserved syntenic regions are highlighted by yellow-red–colored regions***. Putative **orthologs** between barley and the model genomes are connected with ***lines (colored according to model genome chromosomes) between the second and third circles***. Colored lines in the center represent putative **paralogous relationships** between barley chromosomes on the basis of fl-cDNA supported genes included in the genome zipper models of the seven barley chromosomes.



Barley-Centered Four-Genome Comparative View of Grass Genome Collinearity. The **seven barley chromosomes (Hv1 to Hv7)** are depicted by the inner circle of colored bars exactly as in Figure 1. The ***heat map attached to each chromosome indicates the density of barley fl-cDNAs anchored and positioned along the chromosomes according to the genome zipper models.*** Gene density is colored according to the heat map scale. Moving outwards, the bars represent a schematic diagram of the barley chromosomes colored according to **conserved synteny** with the genomes of Brachypodium (Bd), rice (Os), and sorghum (Sb), respectively. In each case, the chromosome numbers and segments are colored according to the chromosome color code (i.e., chr1 through chr5 for Bd, chr1 through chr12 for Os, and chr1 through chr10 for Sb). As in Figure 1, boxes extending from the colored bars indicate structural changes (e.g., inversions) between the gene order in barley and the respective model genome. To the outside of each model genome chromosome, box graphs show the z-score derived from a sliding window analysis of the frequency of fl-cDNAs present at a conserved syntenic position with their corresponding orthologs in Bd, Os, and Sb, respectively (see Methods for a full description of the analysis). A z-score >0 indicates higher than the average **conservation of synteny**, and a z-score <0 highlights decreased syntenic conservation. ***The data points in the center of the diagram depict the Ka/Ks ratios between barley full-length genes and their orthologs in Bd, Os, and Sb.*** Values against Bd are plotted as dark red rectangles, against Os in red circles, and against Sb in blue triangles.



The chromosomes of oil palm.

E. guineensis has ***16 chromosome pairs***, ordered by size, which correspond to 16 linkage groups identified by genetic mapping. Tracks displayed are:

a, gene density;

b, methyl-filtered read density;

c, retroelement density;

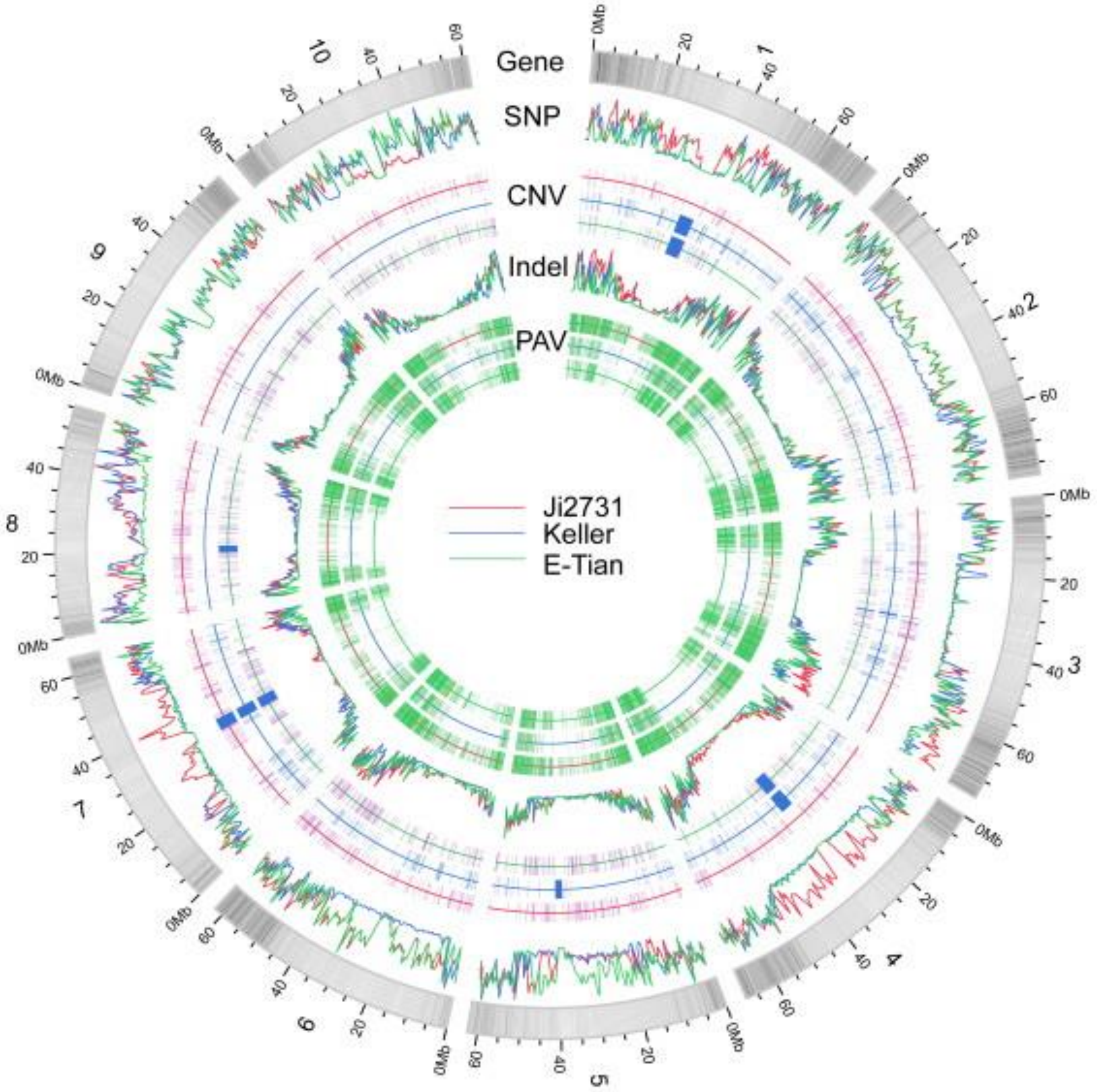
d, simple sequence repeats;

e, low copy number repetitive elements, including telomere repeat TTTAGGG (green), 5S rRNA (orange) and pericentromeric repeats (purple);

f, regional GC content (range 0.3–0.45);

g, genetically mapped scaffolds from the P5-build; and

h, segmental duplications.

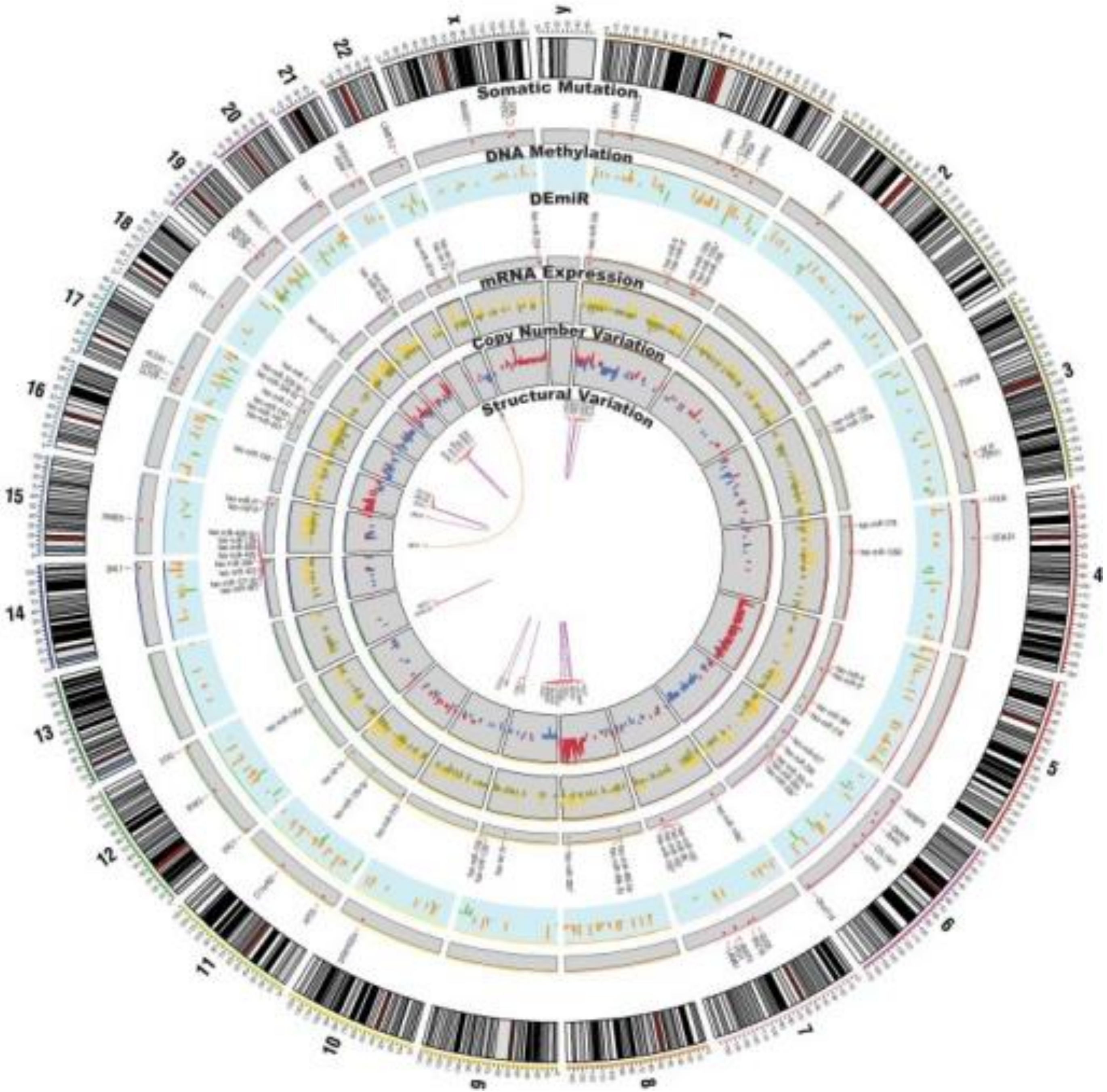


Genome-wide landscape of genetic variation in Sorghum bicolor.

Gene density of chromosomes is visualized by line darkness; the **more genes on a chromosome region, the darker the color**.

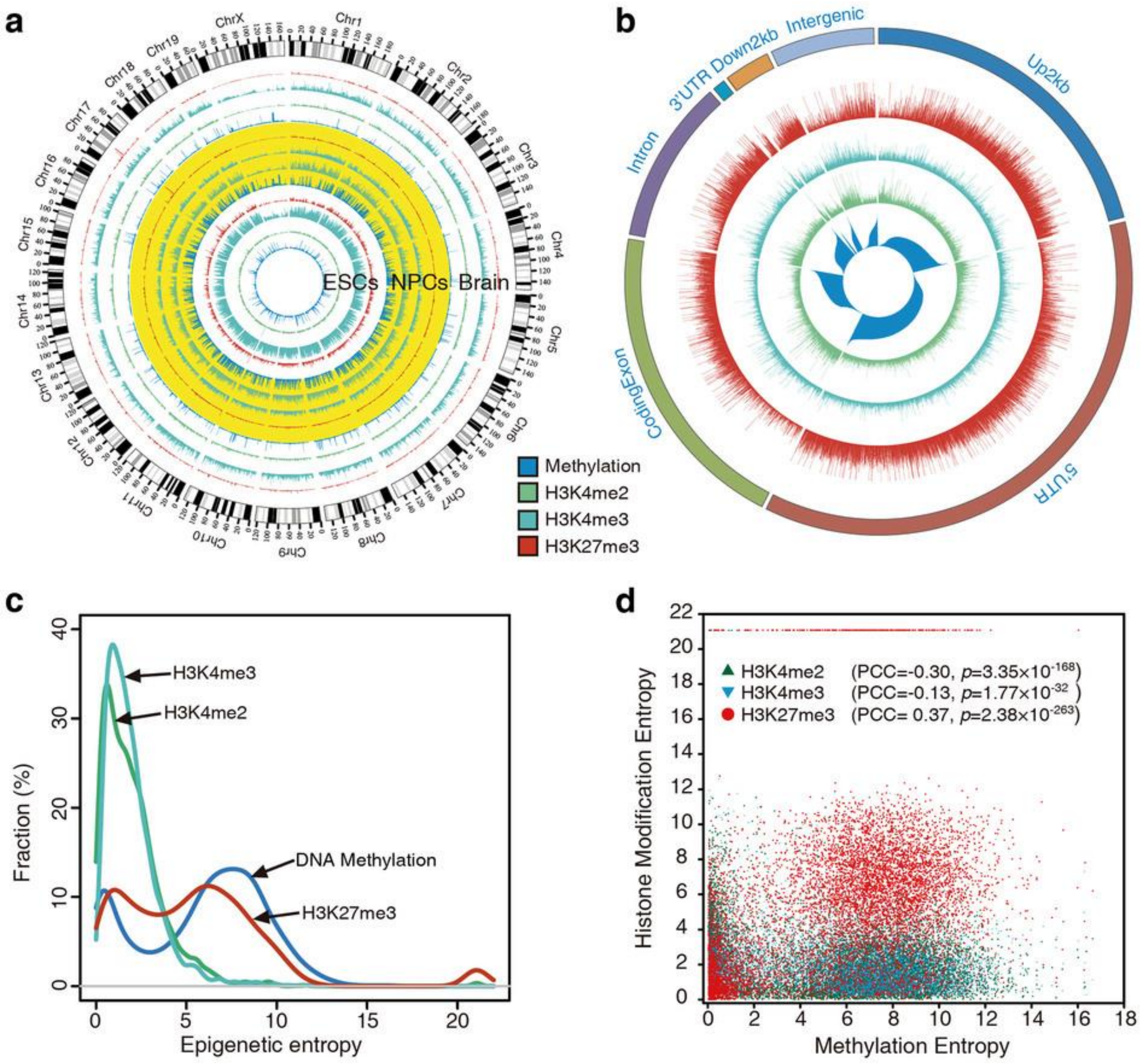
The purple and blue colors in the CNV ring represent gain and loss of copy number variation, respectively.

For PAVs (presence/absence of variation), the green color stands for the absence of variation, whereas pink for the presence of variation.



A High-Dimensional, Deep-Sequencing Study of Lung Adenocarcinoma in Female Never-Smokers. Our study not only provides an overview of the ***alterations occurring in lung adenocarcinoma at multiple levels from genome to transcriptome and epigenome***, but also offers a model for integrative genomics analysis and proposes potential target pathways for the control of lung adenocarcinoma.

Circos plot of somatic mutations, copy number variations, transcriptome expression, and structural variations. From inside to out, structural variations (purple and orange), copy number variations (gain in dark red, loss in dark blue, mRNA expression (up in gold, down in olive), differentially expressed microRNAs (up in red, down in green), DNA methylation with sky-blue background (up in dark orange, down in chartreuse), somatic mutations with a gene symbols, and chromosomal cytobands.



Dynamic epigenetic modifications in CpG islands (CGIs).

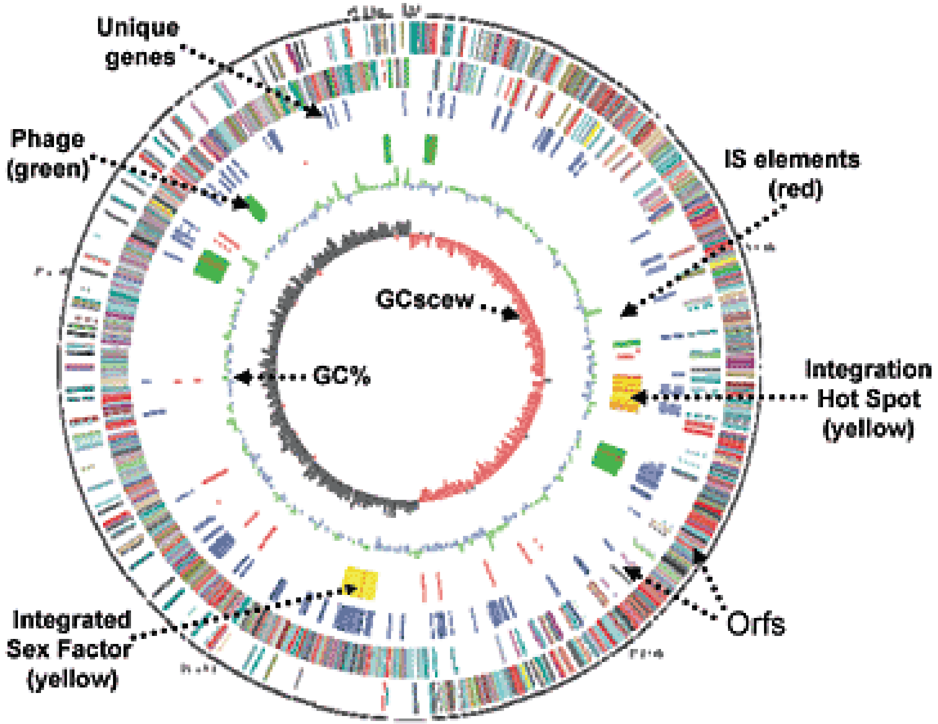
(a) **Circos plot** of the ***epigenetic modification profiles for CGIs*** in the whole genome. The tracks, from outermost to innermost, show the ideogram for the ***mouse karyotype*** (using genome build mm8), and the ***four epigenetic modifications in brain, NPCs (neural progenitor cells) and ESCs***. The tracks are scaled separately to show modification fluctuations.

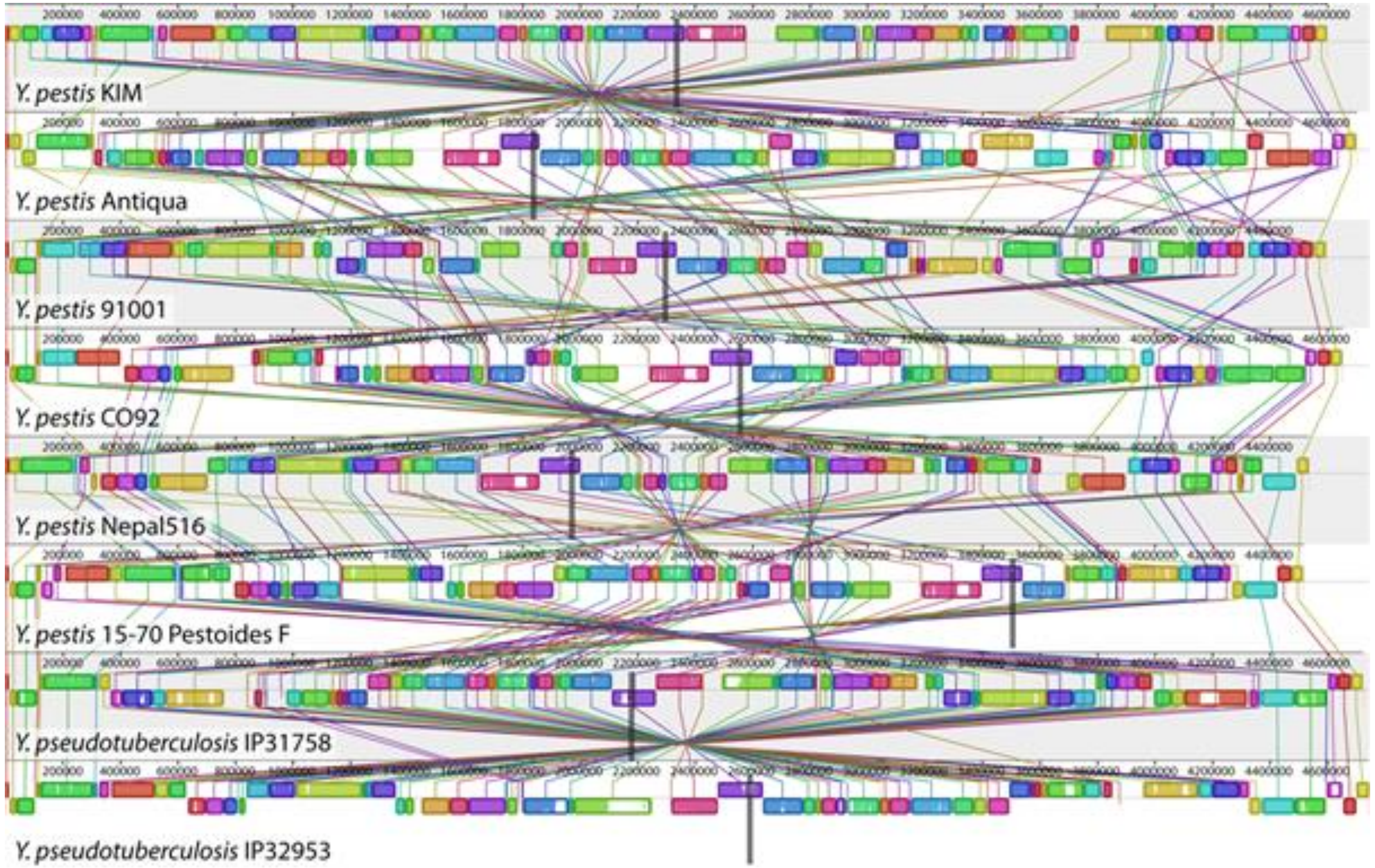
(b) ***Circos plot of the entropy of the four kinds of epigenetic modifications in the different genomic regions***. The tracks, from outermost to innermost, show the genome region, H3K27me3, H3K4me3, H3K4me2 and DNA methylation.

(c) ***Distribution of the epigenetic entropies*** representing the ***variation of epigenetic modifications during neural differentiation, with lower entropy representing greater epigenetic variation***.

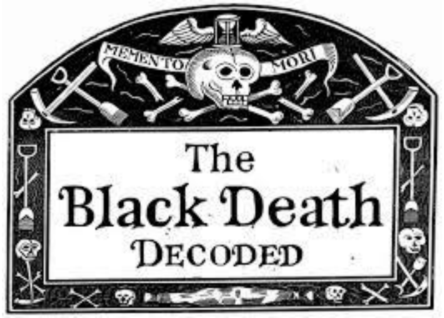
(d) ***Scatter diagram of DNA methylation entropy and the three kinds of histone modification entropy***. PCC is the Pearson correlation coefficient between DNA methylation entropy and one kind of histone modification entropy; p is the significance of the PCC.

Genome atlas of the chromosome of Lactococcus lactis MG1363





Y. pestis KIM
Y. pestis Antiqua
Y. pestis 91001
Y. pestis CO92
Y. pestis Nepal516
Y. pestis 15-70 Pestoides F
Y. pseudotuberculosis IP31758
Y. pseudotuberculosis IP32953



Dynamics of Genome Rearrangement in Bacterial Populations: A genome alignment of eight Yersinia isolates.

Genome structure variation has profound impacts on phenotype in organisms ranging from microbes to humans, yet little is known about how natural selection acts on genome arrangement.

Pathogenic bacteria such as Yersinia pestis, which causes bubonic and pneumonic plague, ***often exhibit a high degree of genomic rearrangement***.

The recent availability of several Yersinia genomes offers an unprecedented opportunity to study the ***evolution of genome structure and arrangement.***

Whole-genome sequencing has revealed that organisms exhibit **extreme variability in chromosome structure**. One common type of chromosome structure variation is **genome arrangement variation: changes in the ordering of genes on the chromosome**.

Not only do we find differences in genome arrangement across species, but in some organisms**, members of the same species have radically different genome arrangements**.

Yersinia exhibit substantial variation in genome arrangement both within and across species.

We discovered an ***excess of rearrangement activity near the origin of chromosomal replication*** and found evidence for a preferred configuration for the relative orientations of the origin and terminus of replication.

Genome-wide intraspecific DNA-sequence variations in rice:

extensive microcolinearity in gene order and content. However, deviations from colinearity are frequent owing to **insertions or deletions**. Intraspecific sequence polymorphisms commonly occur in both coding and non-coding regions. These variations often affect gene structures and may contribute to intraspecific phenotypic adaptations.

Sequence comparison of an orthologous region (of about 100 Kb) from the two cultivated rice subspecies, Oryza sativa L. ssp indica (cv. GLA4) and Oryza sativa L. ssp japonica (cv. Nipponbare).

Light-gray shading indicates their homologous regions and **black bars show the insertions or deletions (Indels) that have occurred in the two subspecies**. Repetitive elements are shown by bars of different colors. Predicted gene orders and structures in the top (Gene +) and bottom (Gene -) strands are indicated in dark blue and red, respectively. MITE, miniature inverted-repeat transposable element.

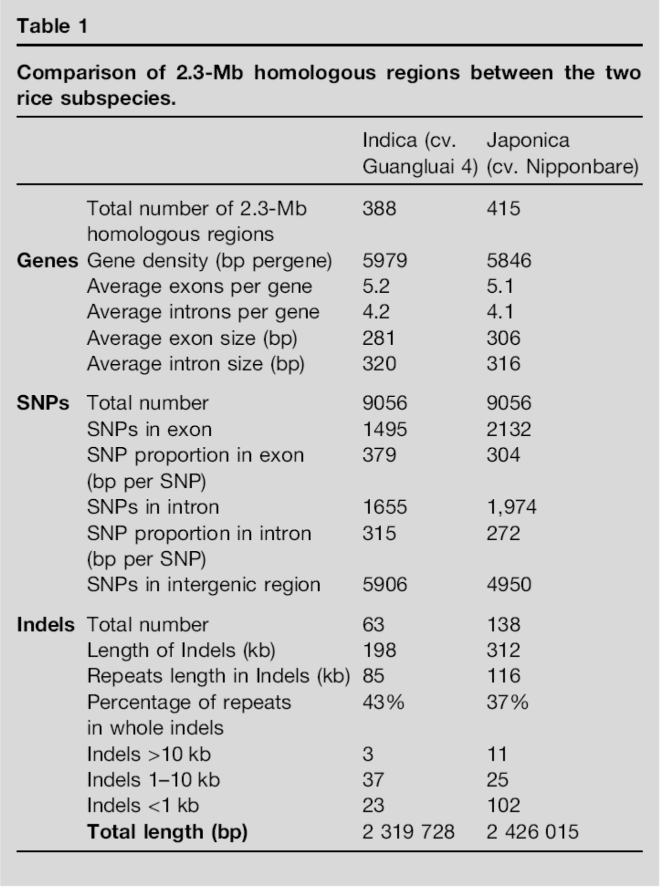
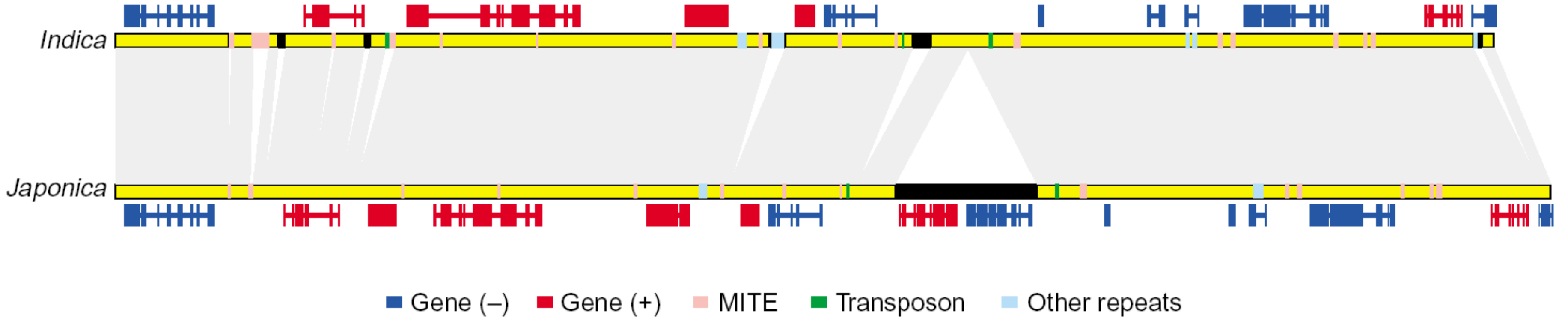


Table 1

Comparison of 2.3-Mb homologous regions between the two rice subspecies.

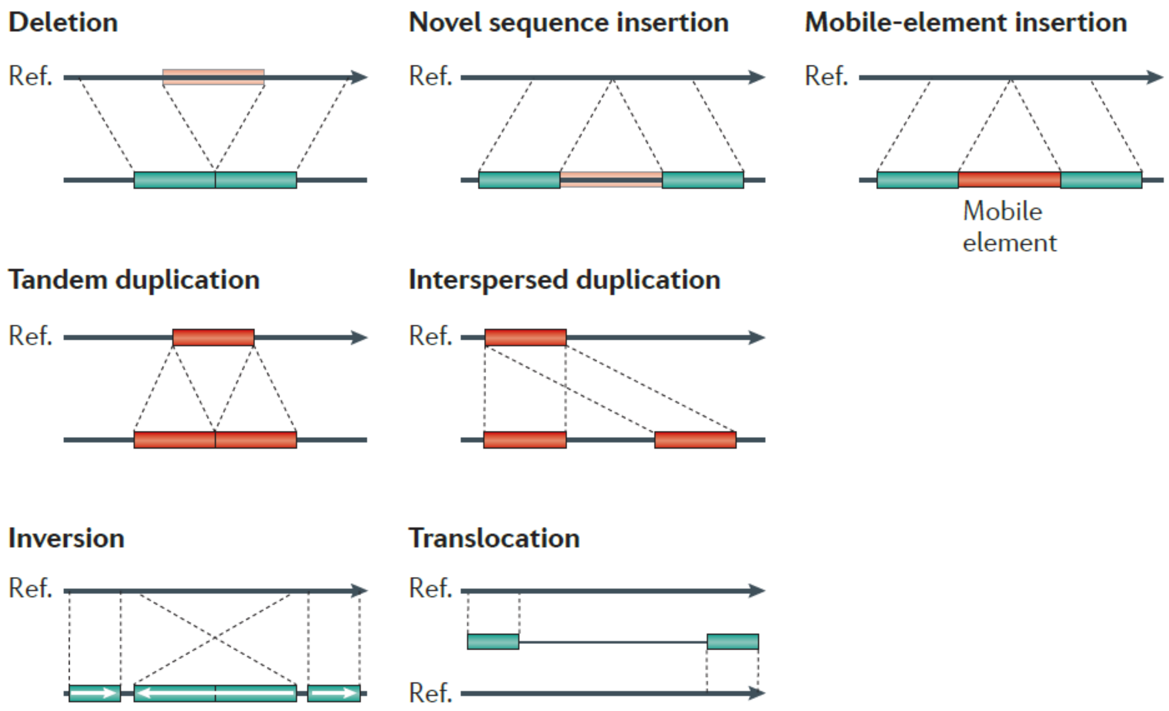
		Indica (cv. Guangluai 4)	Japonica (cv. Nipponbare)
	Total number of 2.3-Mb homologous regions	388	415
Genes	Gene density (bp pergene)	5979	5846
	Average exons per gene	5.2	5.1
	Average introns per gene	4.2	4.1
	Average exon size (bp)	281	306
	Average intron size (bp)	320	316
SNPs	Total number	9056	9056
	SNPs in exon	1495	2132
	SNP proportion in exon (bp per SNP)	379	304
	SNPs in intron	1655	1,974
	SNP proportion in intron (bp per SNP)	315	272
	SNPs in intergenic region	5906	4950
Indels	Total number	63	138
	Length of Indels (kb)	198	312
	Repeats length in Indels (kb)	85	116
	Percentage of repeats in whole indels	43%	37%
	Indels >10 kb	3	11
	Indels 1–10 kb	37	25
	Indels <1 kb	23	102
	Total length (bp)	2 319 728	2 426 015

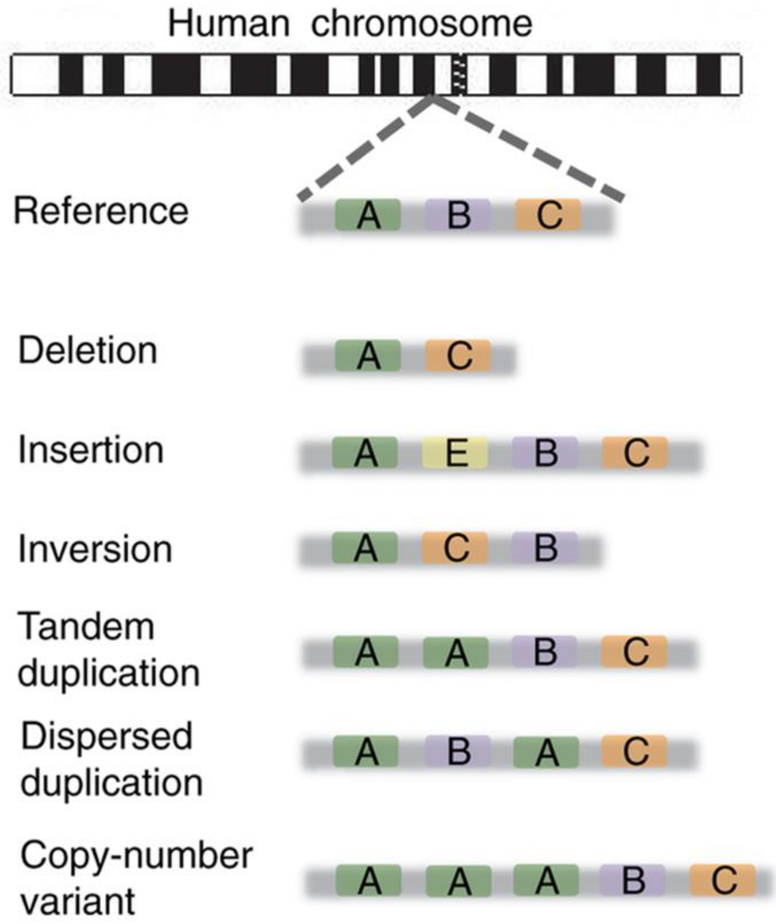


Current Opinion in Plant Biology

Structural variation in the genome: Segmental Duplications (SD) and Copy Number Variations (CNV)

Classes of structural variation





Classes of structural variation. Traditionally, structural variation refers to genomic alterations that are larger than 1 kb in length, but advances in discovery techniques have led to the detection of smaller events. Currently, **>50 bp is used as an operational demarcation between indels and copy number variants (CNVs)**.

The schematic depicts deletions, novel sequence insertions, mobile-element insertions, tandem and interspersed segmental duplications, inversions and translocations in a test genome (lower line) when compared with the reference genome.

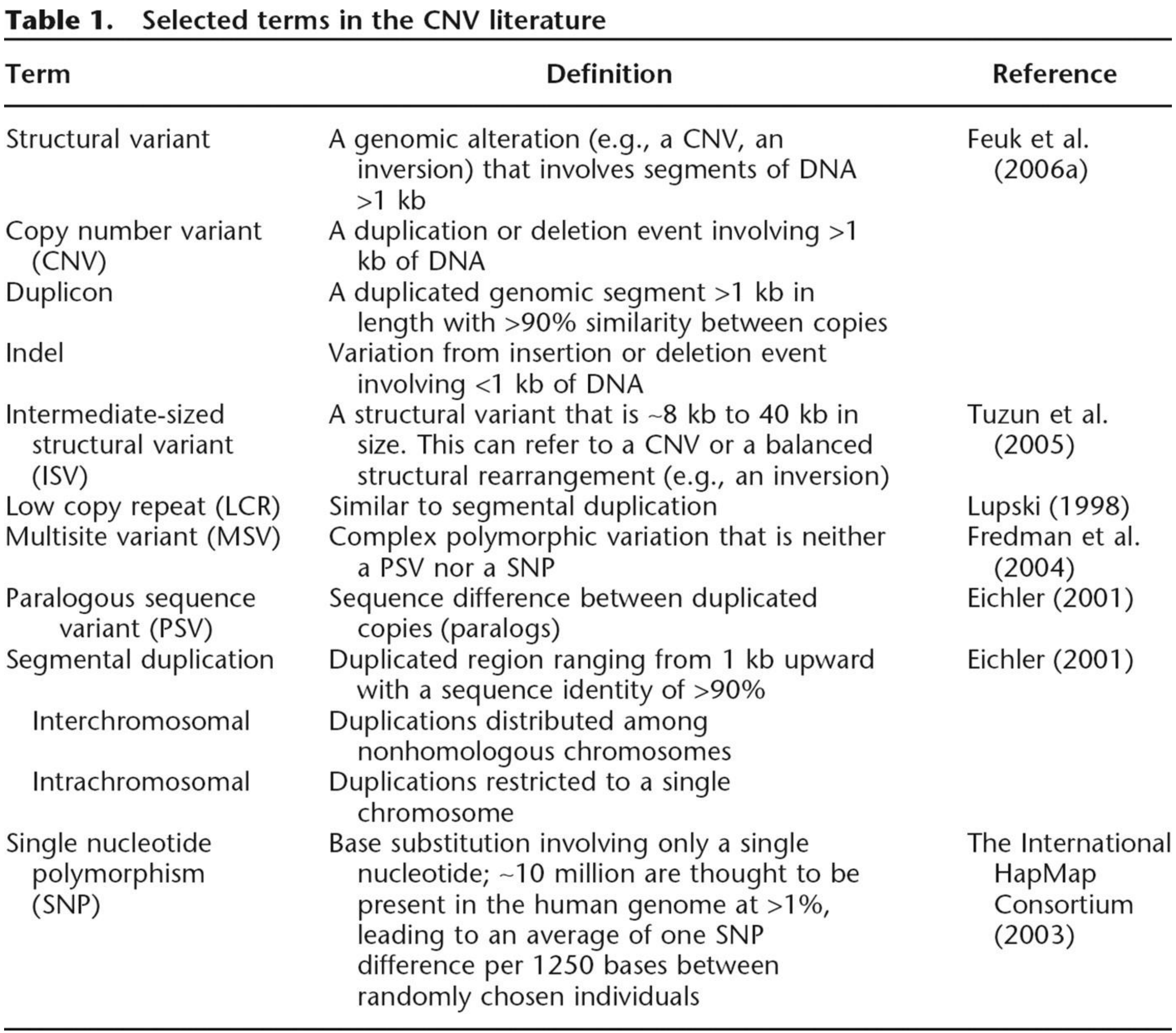
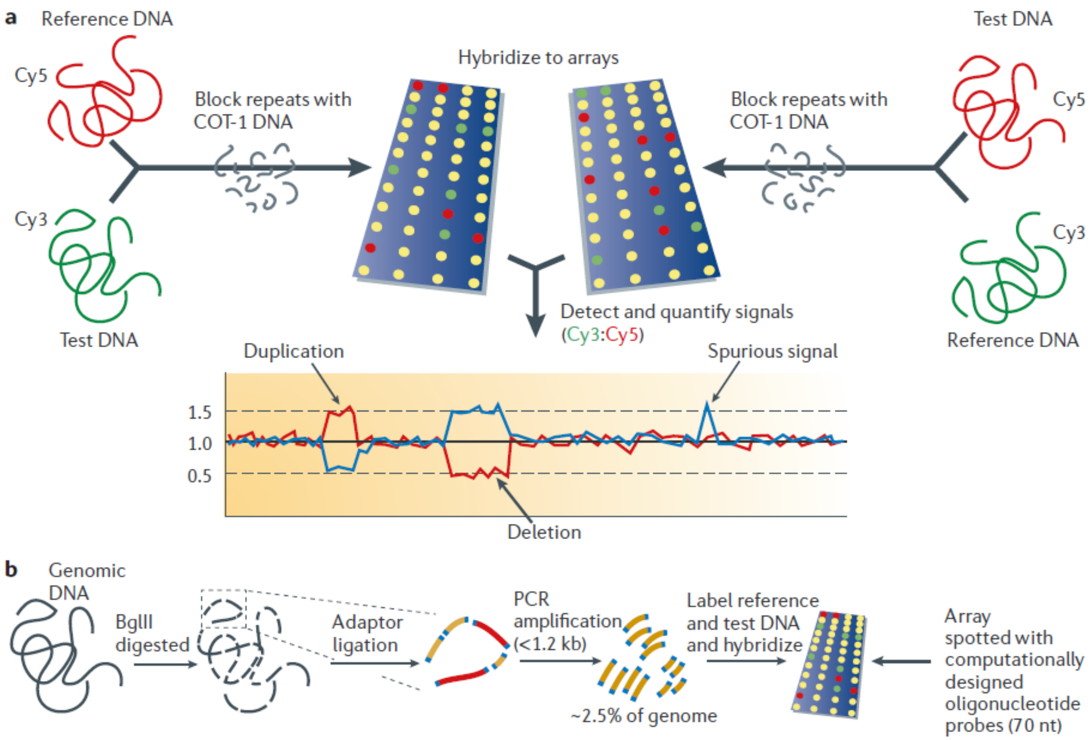


Table 1. Selected terms in the CNV literature

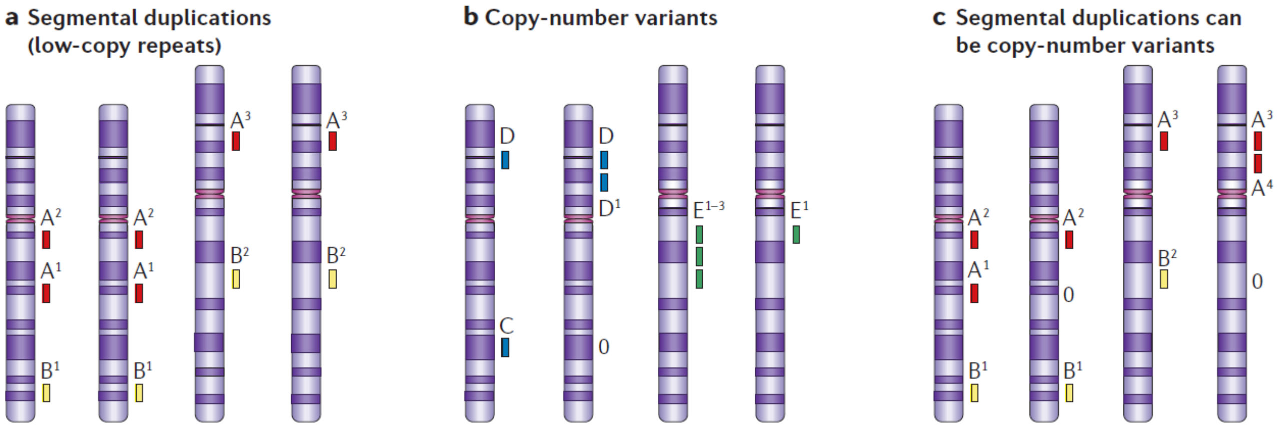
Term	Definition	Reference
Structural variant	A genomic alteration (e.g., a CNV, an inversion) that involves segments of DNA >1 kb	Feuk et al. (2006a)
Copy number variant (CNV)	A duplication or deletion event involving >1 kb of DNA	
Duplicon	A duplicated genomic segment >1 kb in length with >90% similarity between copies	
Indel	Variation from insertion or deletion event involving <1 kb of DNA	
Intermediate-sized structural variant (ISV)	A structural variant that is ~8 kb to 40 kb in size. This can refer to a CNV or a balanced structural rearrangement (e.g., an inversion)	Tuzun et al. (2005)
Low copy repeat (LCR)	Similar to segmental duplication	Lupski (1998)
Multisite variant (MSV)	Complex polymorphic variation that is neither a PSV nor a SNP	Fredman et al. (2004)
Paralogous sequence variant (PSV)	Sequence difference between duplicated copies (paralogs)	Eichler (2001)
Segmental duplication	Duplicated region ranging from 1 kb upward with a sequence identity of >90%	Eichler (2001)
Interchromosomal	Duplications distributed among nonhomologous chromosomes	
Intrachromosomal	Duplications restricted to a single chromosome	
Single nucleotide polymorphism (SNP)	Base substitution involving only a single nucleotide; ~10 million are thought to be present in the human genome at >1%, leading to an average of one SNP difference per 1250 bases between randomly chosen individuals	The International HapMap Consortium (2003)

Structural variation in the human genome

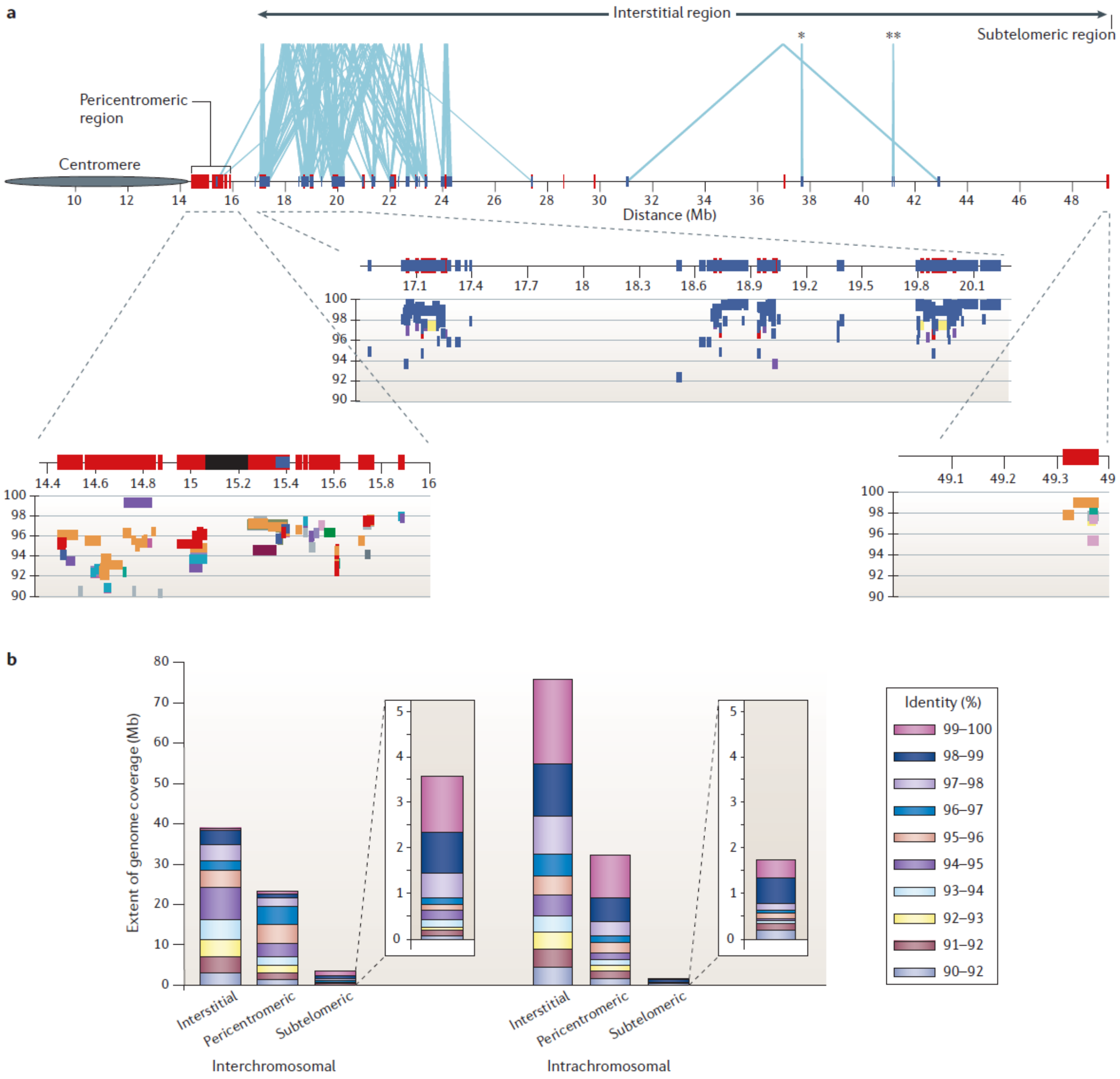

a
Reference DNA
Test DNA
Cy5
Cy3
Block repeats with COT-1 DNA
Hybridize to arrays
Block repeats with COT-1 DNA
Cy5
Cy3
Test DNA
Reference DNA
Detect and quantify signals (Cy3:Cy5)
Duplication
Spurious signal
1.5
1.0
0.5
Deletion
b
Genomic DNA
BglII digested
Adaptor ligation
PCR amplification (<1.2 kb)
~2.5% of genome
Label reference and test DNA and hybridize
Array spotted with computationally designed oligonucleotide probes (70 nt)

Array-based, genome-wide methods for the identification of copy-number variants

The complexity of segmental duplications and copy-number variants

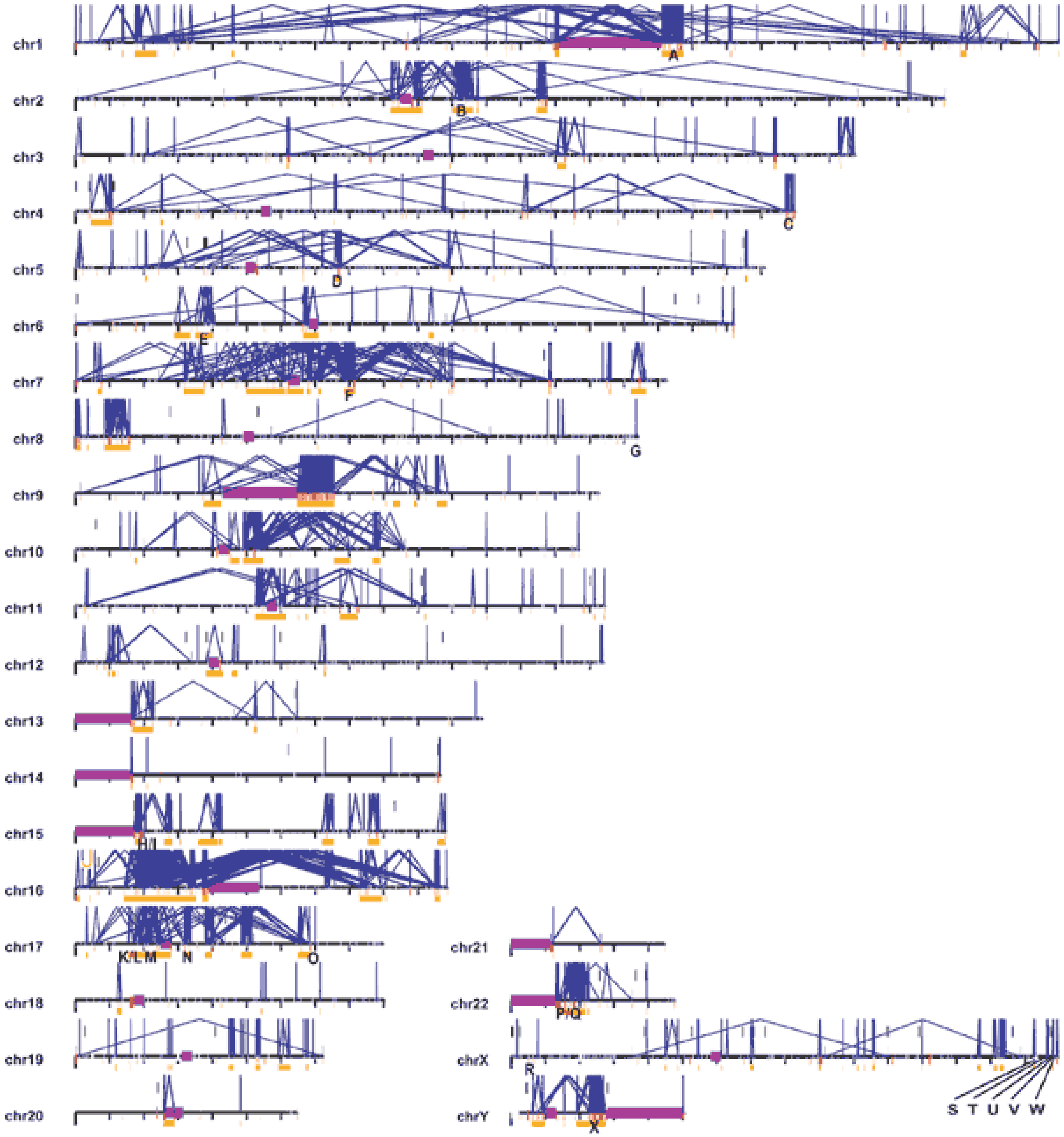

a Segmental duplications (low-copy repeats)
A2
A1
B1
A2
A1
B1
A3
B2
A3
B2
b Copy-number variants
D
C
D
D1
0
E1–3
E1
c Segmental duplications can be copy-number variants
A2
A1
B1
A2
0
B1
A3
B2
A3
A4
0

The distribution of segmental duplications (SDs) in the human genome



a| The q-arm of chromosome 22 is used to show the distribution of the three classes of SD: pericentromeric, interstitial and subtelomeric. The overview (top panel) shows the position of SDs that are greater than or equal to 10 kb in length and greater than or equal to 90% identity. Interchromosomal and intrachromosomal pairwise SDs are shown in red and blue, respectively, with light blue lines joining homologous SDs. The interstitial region shows examples of interspersed (*) and tandem (**) intrachromosomal duplications. The subtelomeric region contains approximately 100 kb of interchromosomal SD sequence, which is shared with up to three other non-homologous subtelomeric regions.

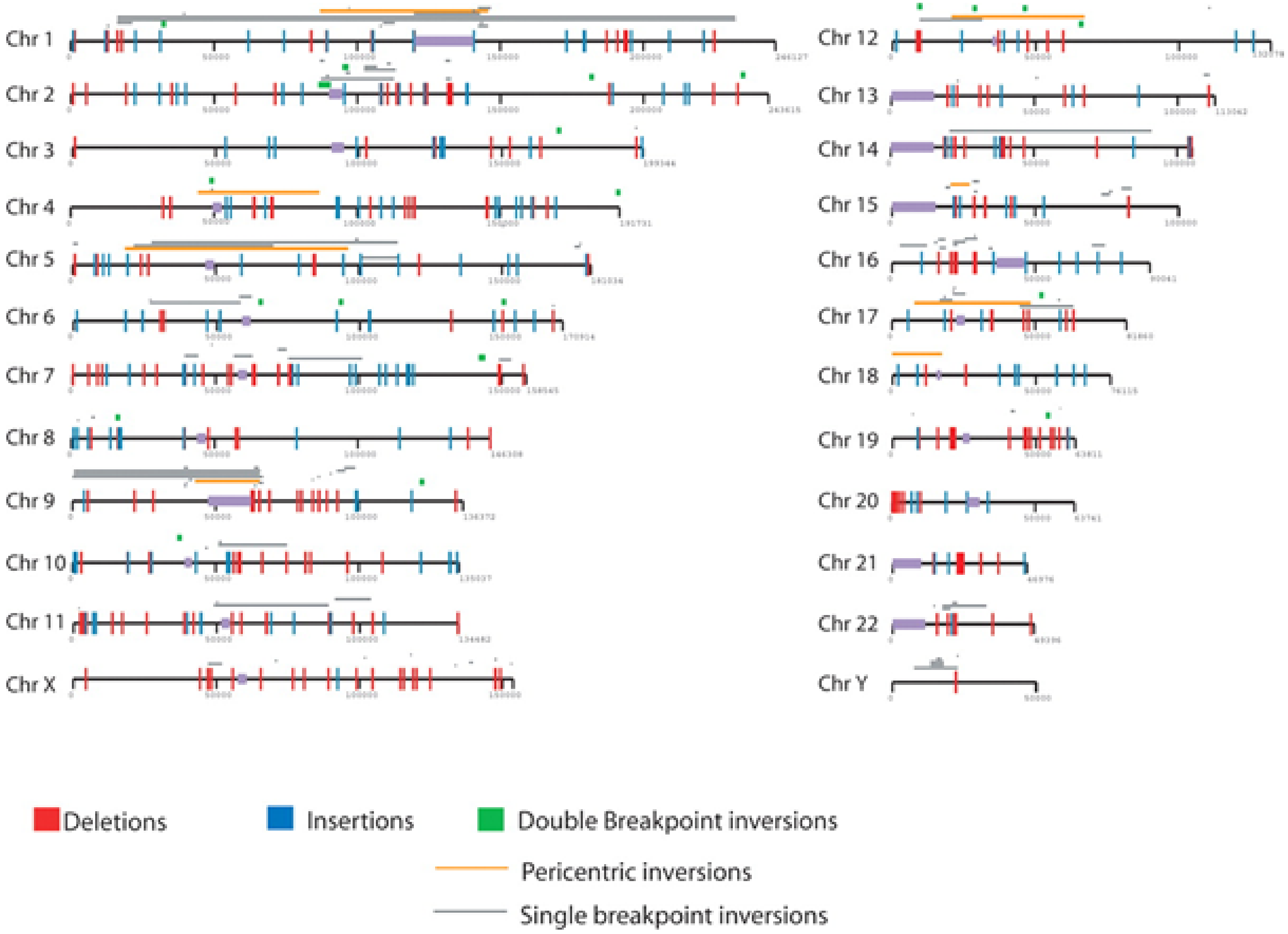
b| The distribution of interchromosomal and intrachromosomal duplications within the pericentromeric, subtelomeric and interstitial regions are shown in terms of % identity of alignments. The y-axis represents the total number of non-redundant base pairs within the genome (build 35) that consist of each type of SD. The distribution within each category was calculated as the proportion of pairwise alignments at each % identity.



Genome-wide view of duplication.

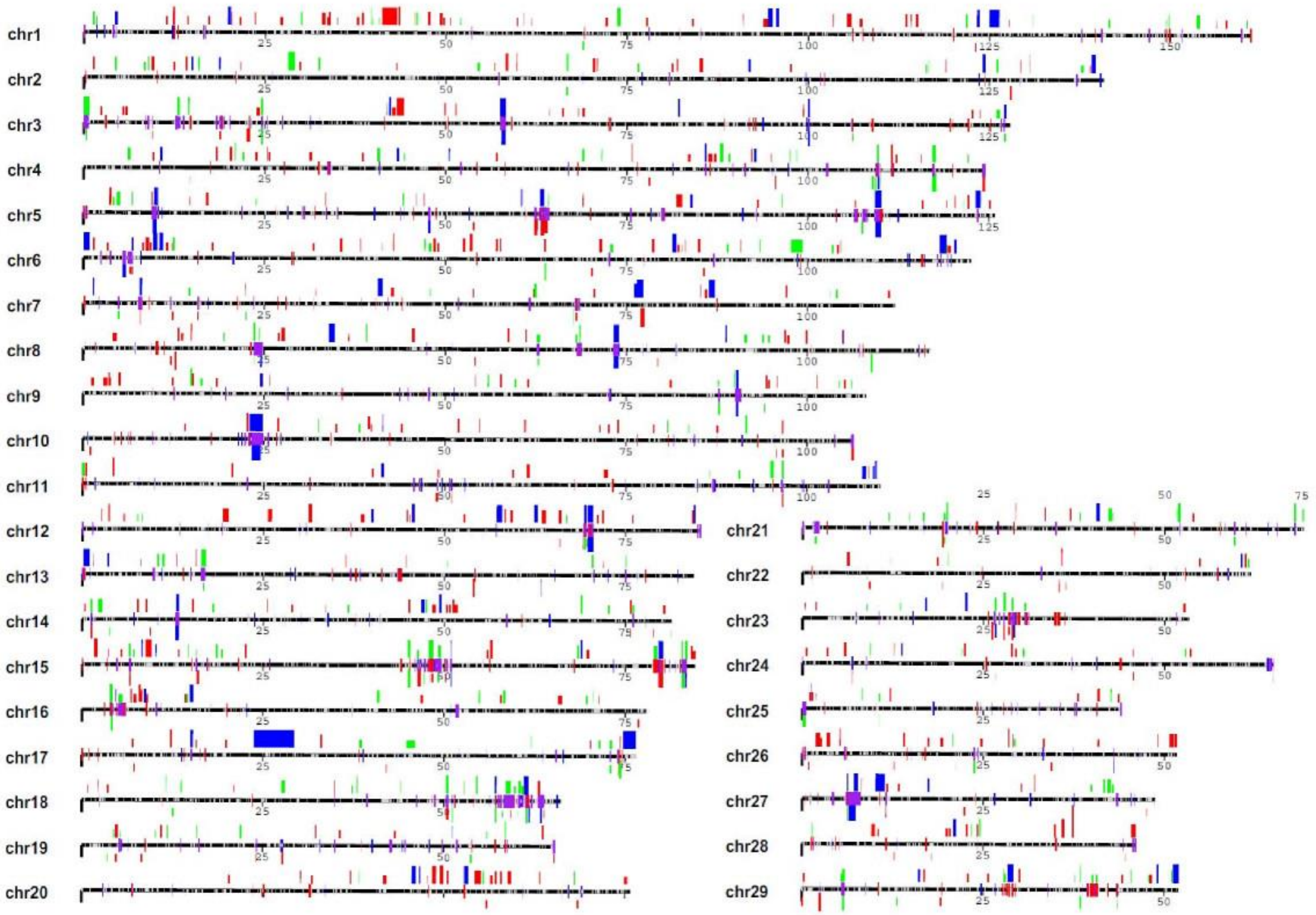
Gold bars indicate "hot spots of genomic instability."

Patterns of intrachromosomal and interchromosomal duplication (≥10 kb; ≥95%). The graphic shows a genome-wide view of **intrachromosomal (blue**, with connecting lines) and **interchromosomal (red** bars) **segmental duplications**. Purple bars represent areas (acrocentric chromosomal arms, heterochromatin satellite DNA, and centromeres) not targeted as part of the Human Genome Project. ***Unique regions (≥50 kb and <10 Mb) of the genome encompassed by intrachromosomal duplications (≥95% sequence identity and ≥10 kb) are shown as gold bars***. *Such regions are typically associated with recurrent chromosomal structural rearrangements associated with genetic disease. A total of 169 regions (~298 Mb of sequence) were identified as potential hot spots for genomic rearrangement. Twenty-four of these regions (labeled A to X) correspond to known genomic disorders.*



Summary of structural variation between chimpanzee and human.

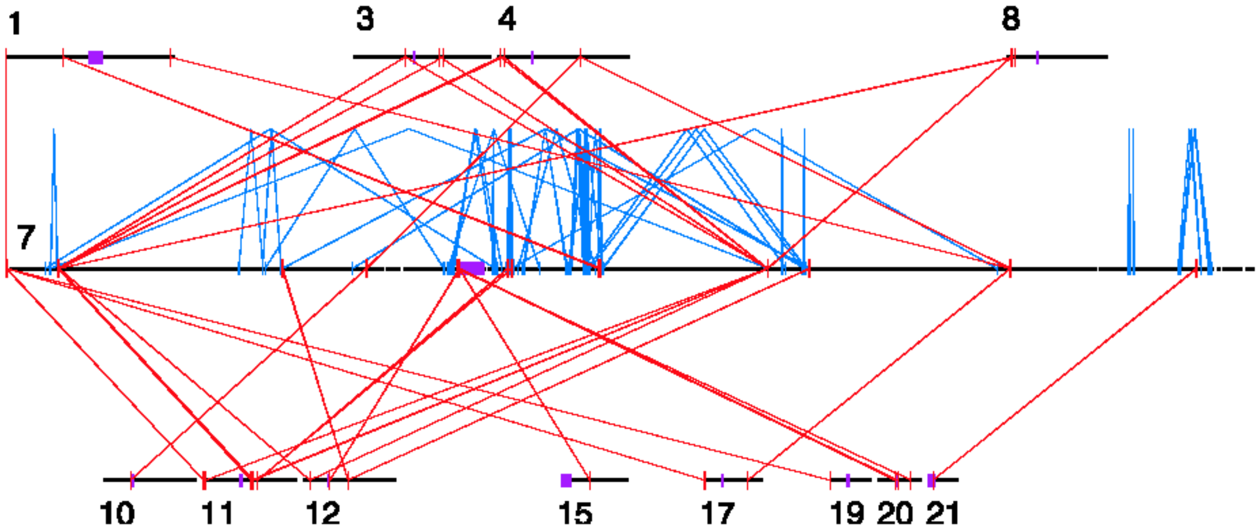
A diagram of the location of ***all 651 structural variants between humans and chimpanzee*** **mapped to the human reference assembly**. Chimpanzee deletions (n = 293) are shown in red; insertions (n = 184) are shown in blue. Inversions/duplicative transpositions (n = 174) are classified into three groups: confirmed ***pericentric*** cytogenetic inversions from Yunis and Prakash (1982) (orange); ***double breakpoint inversions***, if both of the breakpoints were captured (green); and ***single breakpoint inversions***, if only one end was captured (gray). A significant fraction of the latter corresponds to duplicative transpositions of segmental duplications as opposed to bona fide inversions.



Genomic landscape of cattle copy number variations and segmental duplications.

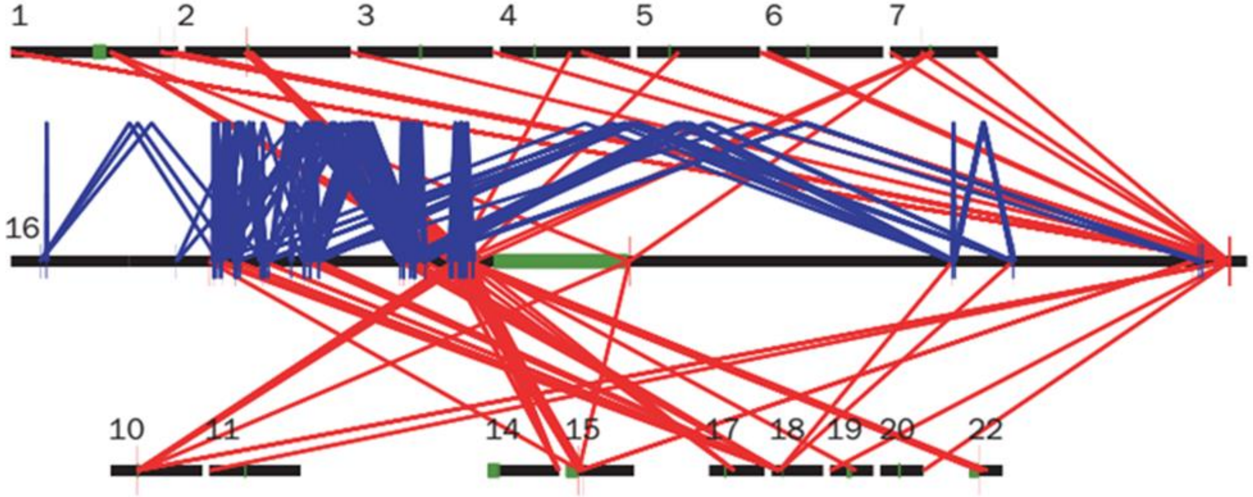
CNV regions (682 events, 139 Mb, ~4.60% of the bovine genome) reported by 521 SNP genotyped individuals are shown above the chromosomes in ***green (gain), red (loss) and dark blue (both),*** while below are the CNV regions (177 events, 28 Mb, ~1% of the bovine genome) reported by 90 array CGH experiments by Liu et al. The bar height represents their frequencies: short (appeared in 1 sample), median (≥2 samples) and tall (≥5 samples).

Segmental duplications (94.4 Mb, 3.1% of the bovine genome) predicted by two independent computational approaches are illustrated on the chromosomes in red (WSSD), blue (WGAC) or purple (both). The patterns are depicted for all duplications for ≥5 kb in length and ≥90% sequence identity. The gaps in the assembly are represented on the chromosomes as white ticks.



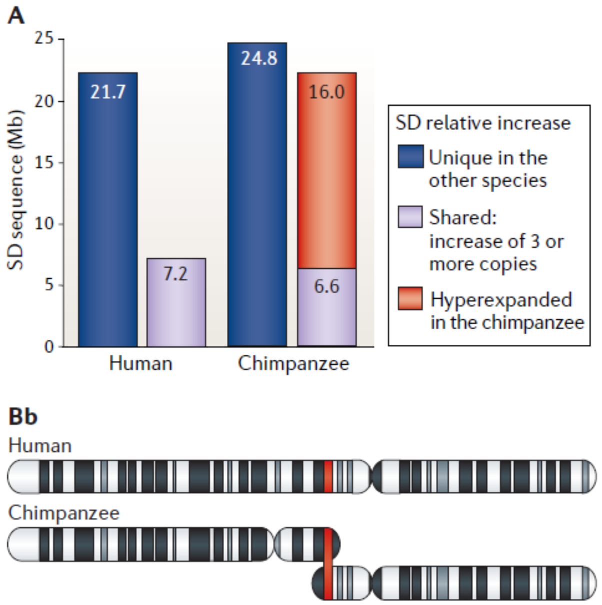
Recent segmental duplications on human chromosome 7

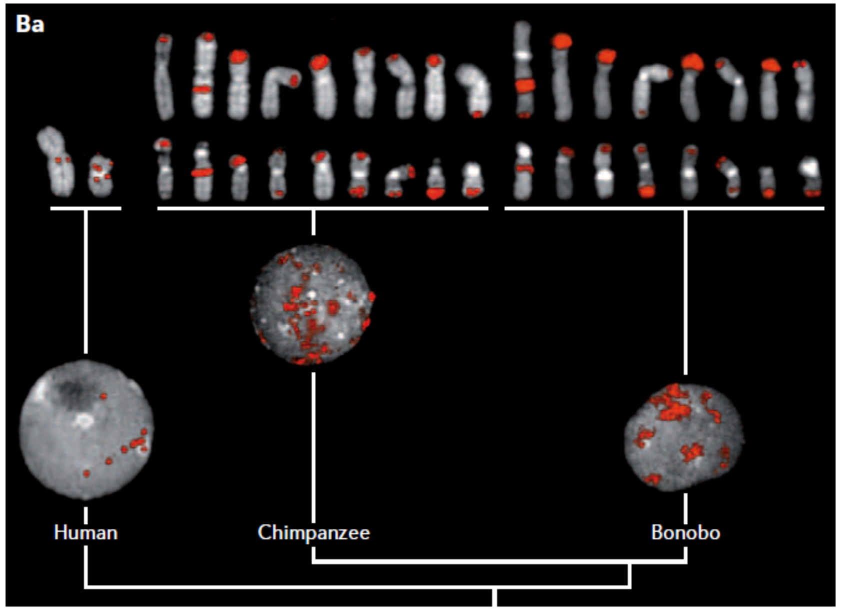
The distribution of both **interchromosomal (red) and intrachromosomal (blue)** duplications is shown for human chromosome 7 (occurred over the last ~30 million years).



Segmental duplication. The horizontal bar in the center is a linear map of the DNA of **human chromosome 16** (the central green segment represents heterochromatin). The black horizontal bars at the top and bottom represent linear maps of 16 other chromosomes containing large segments that are shared with chromosome 16, with red connecting lines marking the positions of homologous sequences. Intrachromosomal duplications are shown by blue chevrons (^) linking the positions of large duplicated sequences on chromosome 16.

Segmental duplication content of hominoids: hyperexpansions in chimpanzee

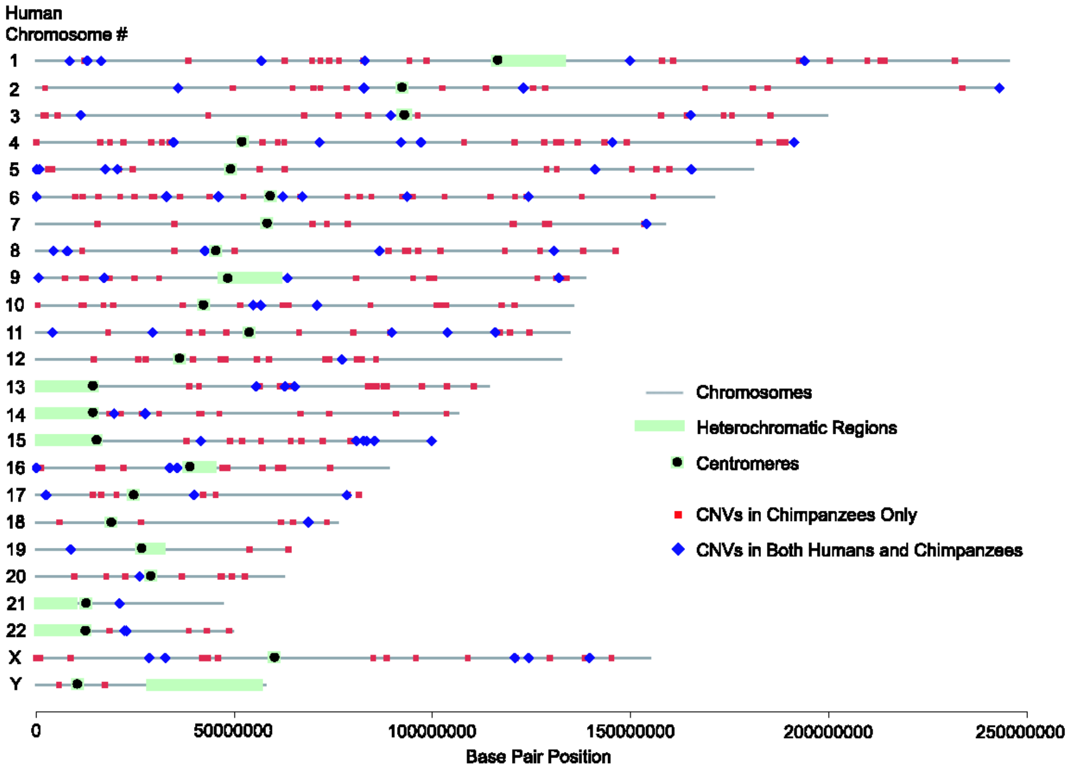




A| A comparison of duplicated sequence from the chimpanzee and human genomes allowed the identification of regions of shared duplication and those that contain human-specific or chimpanzee-specific multicopy sequence (>94% identity and >20 kb). An estimated 60% represented duplicative gain, whereas 40% of the change occurred as a result of deletion of ancestral duplications. **Overall, as shown in the graph, a minimal estimate of 76 Mb is differentially duplicated between humans and chimpanzees, corresponding to 3–5 Mb duplication gain per million years.** This is a conservative estimate owing to limitations of the detectable differences to >94% identity, >20 kb and threefold or greater copy-number difference.

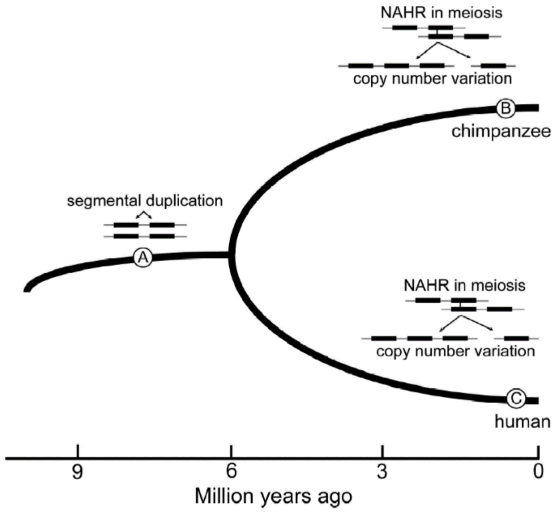
B| Hyperexpansion of an SD in the chimpanzee lineage. Fluorescence in situ hybridization staining (red) is shown in **panel Ba** for an SD duplicon of approx 40 kb, revealing an expansion to 400–500 copies of this sequence in the chimpanzee and bonobo, mainly near telomeres (only chromosomes that carry the duplicon are shown). This expansion added approx 16 Mb of duplicated sequence in a common ancestor of chimpanzees and bonobos that is not present in gorillas and humans. As shown in **panel Bb**, the same SD underlies an association between duplication and rearrangement: it is associated with a large-scale chromosome fusion event that produced human chromosome 2 (this chromosome is shown next to the two orthologous chimpanzee chromosomes; the SD lies 40 kb proximal of the fusion point).

Hotspots for copy number variation in chimpanzees and humans



Copy number variation is surprisingly common among humans and can be involved in **phenotypic diversity and variable susceptibility to complex diseases**.

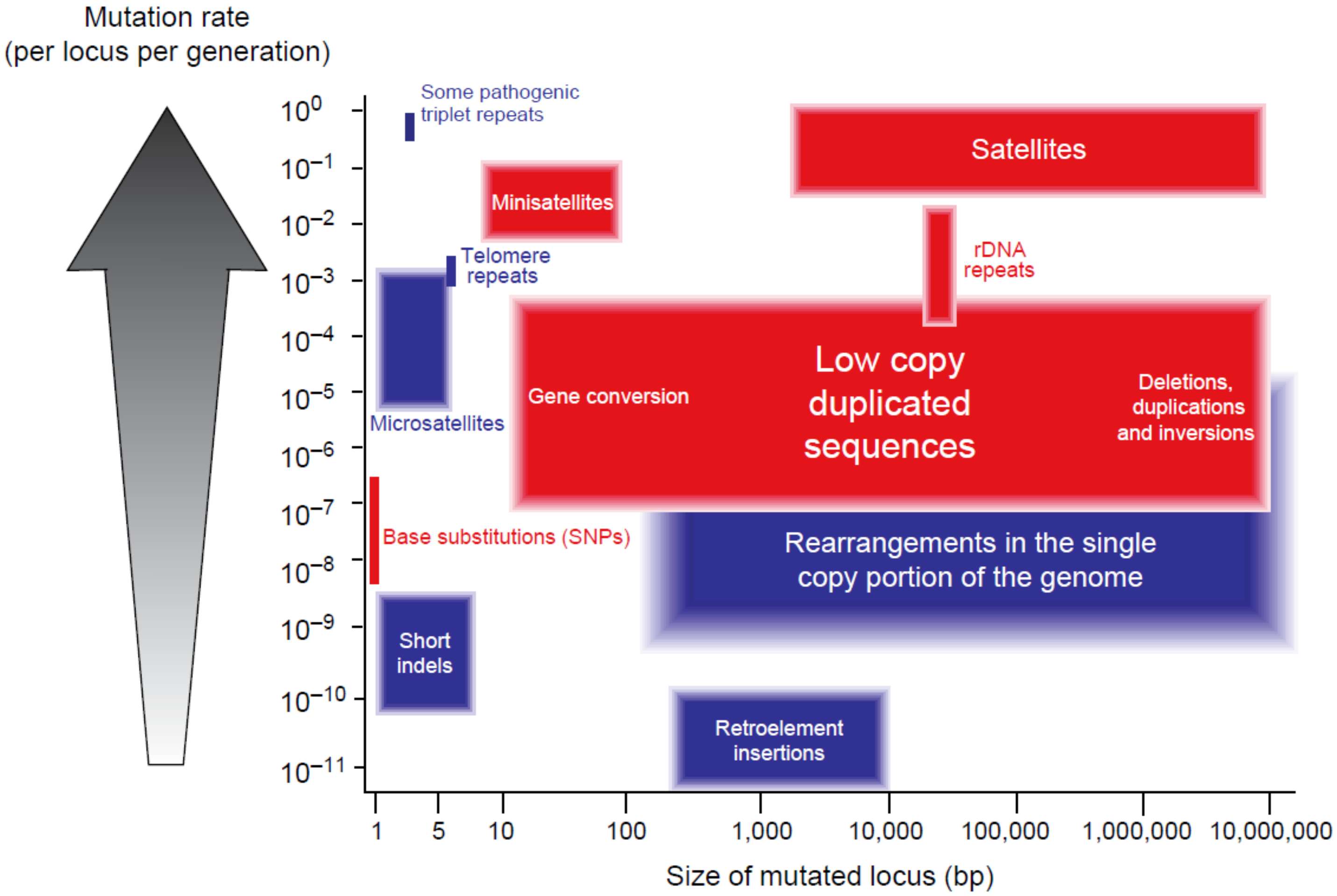
Many CNVs were observed in the corresponding regions in both chimpanzees and humans; especially those CNVs of higher frequency. Strikingly, these loci are enriched 20-fold for **ancestral segmental duplications**, which **may facilitate CNV formation through nonallelic homologous recombination mechanisms**. Therefore, some of these regions may be unstable "hotspots" for the genesis of copy number variation, with recurrent duplications and deletions occurring across and within species.



Model for evolution of CNV hotspots

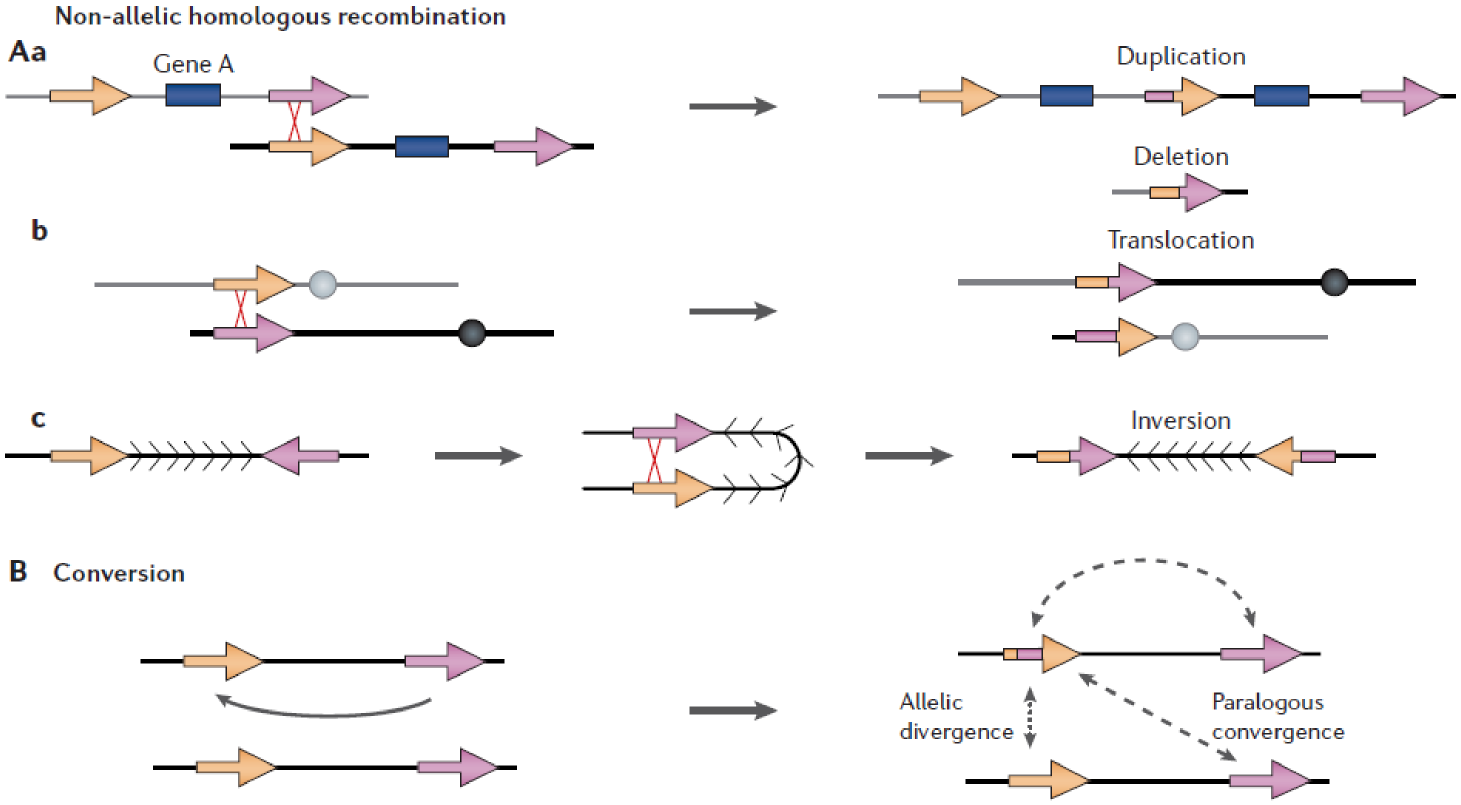
Certain segmental duplications that arose in a human–chimpanzee common ancestor (depicted at point A) may facilitate separate nonallelic homologous recombination (NAHR) in both chimpanzees (B) and humans (C), leading to the genesis of CNVs in both species. If NAHR in these regions occurs frequently, it may be expected to lead to the maintenance of common CNVs by way of recurrent duplications and deletions.

Mutational processes that generates genomic variation



Non-Allelic Homologous Recombination (NAHR) (gene conversion and rearrangement) contribute to sequence variation and structural polymorphism in the human genome.

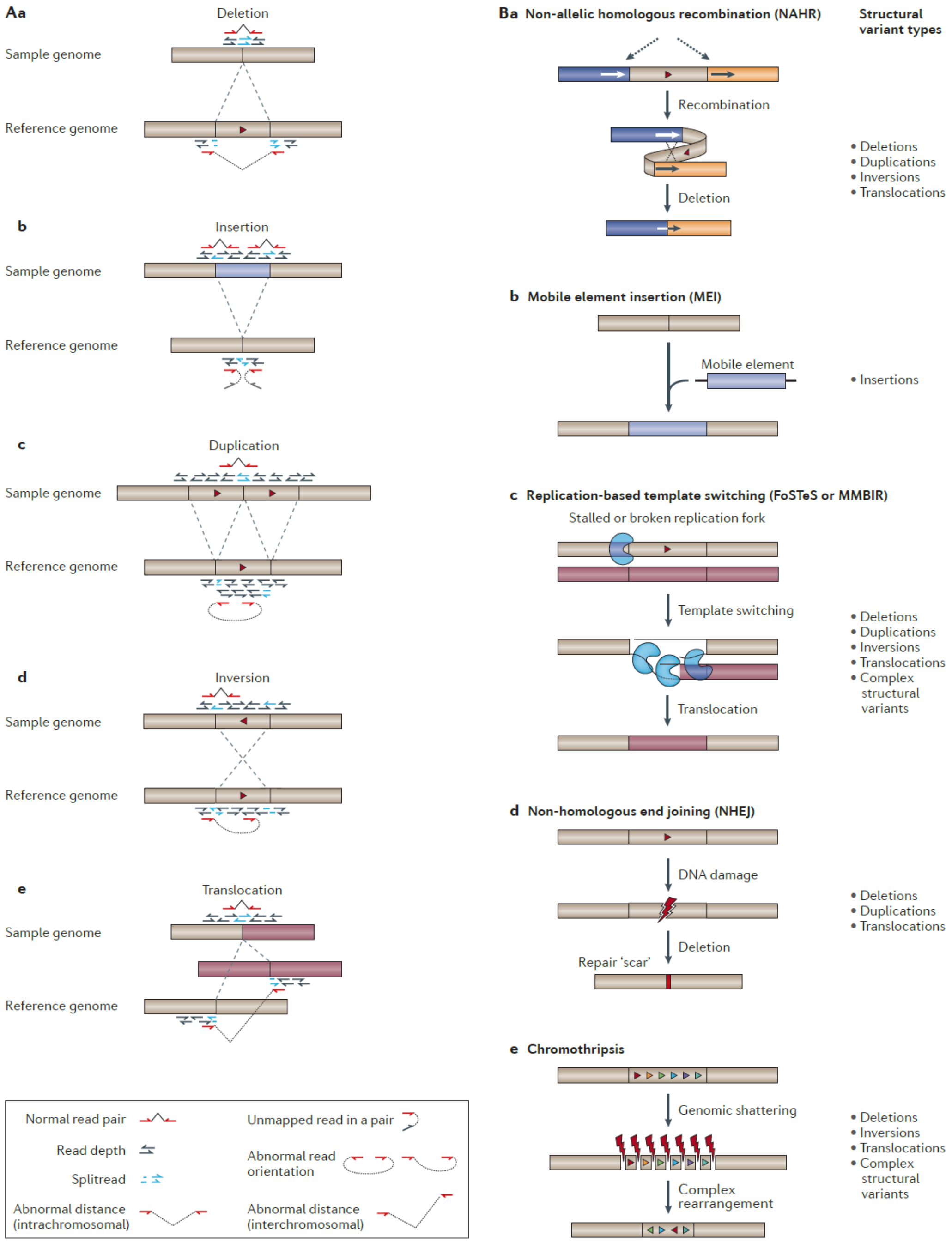
Non-random mechanisms of segmental duplication (SD) evolution



Two major homology-driven forces affect the occurrence and evolution of SDs.

A| ***Non-allelic homologous recombination (NAHR)*** between highly identical SDs causes further rearrangements depending on the location and orientation of the SD copies involved. **Tandem duplications** and intervening deletions can occur as a result of NAHR between adjacent duplicated sequences (Aa). Alternatively, **translocations** can result from exchange between SDs on non-homologous chromosomes (Ab). In both cases, the copies created are in the same orientation as the original SD. By contrast, **inversions** can occur as a consequence of recombination between inverted intrachromosomal duplications (Ac).

B| ***Gene conversion*** between SDs occurs as a result of the transfer of sequence information between copies. Such events can increase allelic diversity at the converted copy and cause homogenization of the SD copies (in a process known as concerted evolution).

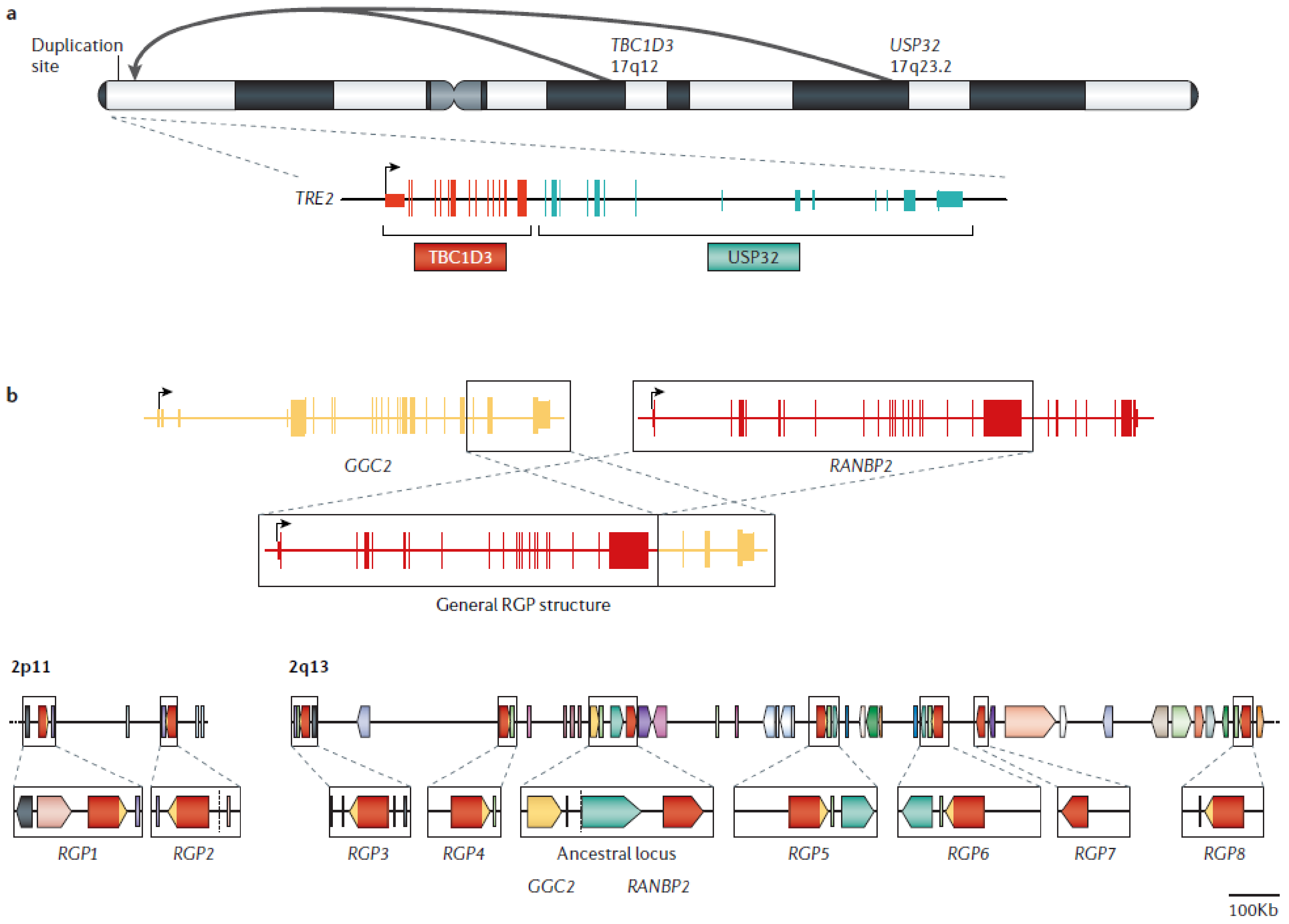


Structural variants: classes and formation mechanisms.

A| Structural variants comprise unbalanced copy-number variations ≥50 bp, including deletions (a), insertions (b) and duplications (c), as well as balanced variants such as inversions (d) and translocations (e). Widely applied DNA-sequencing-based approaches for structural variant detection using the relative orientation, position and read depth of paired-end DNA sequencing reads are indicated. For simplicity, read depth is represented by unpaired reads.

B| Molecular mechanisms leading to structural variant formation. a| Recurrent structural variants often result from non-allelic homologous recombination (NAHR) which involves recombination between long highly similar low-copy-number repeats (blue and orange segments). b| Novel genomic insertions can involve mobile element insertion of transposable elements by retrotransposition. c| DNA-replication-associated template-switching events, involving the fork-stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) mechanisms, can lead to simple or complex structural variants, frequently involving duplicative events. d| Many structural variants in humans are attributed to non-homologous end joining (NHEJ), which is a process that repairs DNA double-strand breaks in the absence of extensive sequence homology and is often accompanied by the addition or deletion of several nucleotides in the form of a 'repair-scar' (small red bar). e| Chromothripsis - which is a phenomenon that seems to involve chromosome shattering leading to numerous breakpoints, followed by error-prone DNA repair — has been proposed to lead to one-off catastrophic rearrangements in several cancers and also in the context of germline DNA rearrangements.

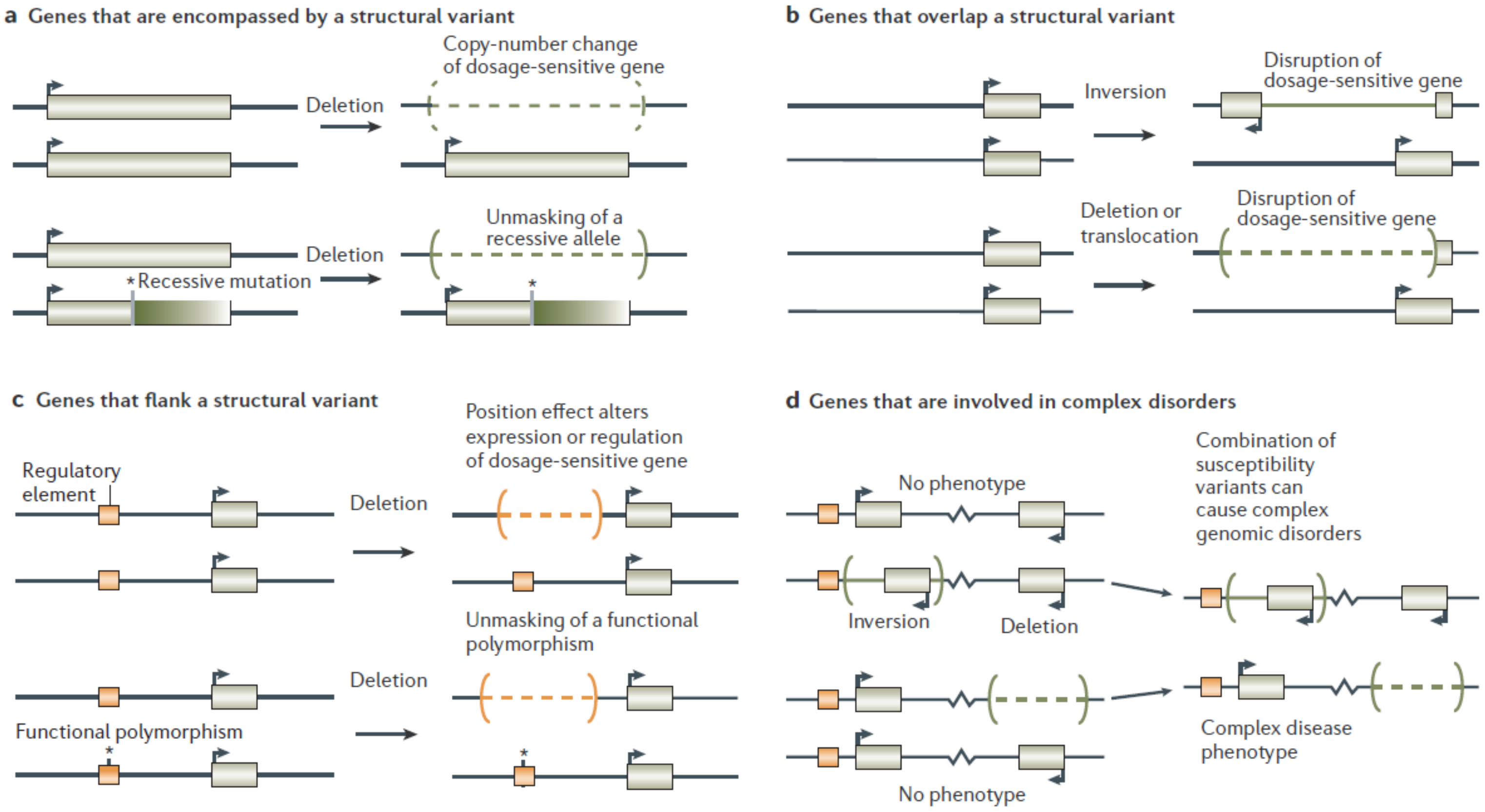
Gene innovation in segmental duplications (SDs)



Two examples of 'novel' primate-specific genes that have been created by SDs are shown.

a| **The TRE2 oncogene (also known as ubiquitin-specific protease 6 (USP 6)) is a hominoid-specific gene that is located at 17p13.2 in humans.** This gene was formed from the fusion of two SDs, each carrying sequence from genes that are located on the q-arm of human chromosome 17, at a distance from the duplication site (TBC1 domain family member 3 (TBC1D3) and USP32). **TRE2 has derived exons 1–14 (red) from TCB1D3 and exons 15–29 (green) from USP32.**

b| The RanBP2-like GRIP-domain-containing protein (RGP) gene family formed from the fusion of SDs of the genes RANBP2 and GGC2, which are located on human chromosome 2q13. As shown in the top panel, a prototypical RGP is composed of the first 20 exons of RANBP2 (red) and last three exons of GGC (yellow). ***This fusion sequence has been extensively duplicated as part of duplication hubs on chromosome 2, on both the q- and p-arms.*** Below this is a detailed view of the ***duplicon structure of the RGP-containing regions involved, illustrating the complex mosaic pattern that has arisen during the formation of this gene family.*** GGC2 and RANBP2 regions are shown in yellow and red, respectively; regions of other genes are shown in various colours. **These duplication hubs show evidence for multiple functional copies under extensive positive selection.**


a Genes that are encompassed by a structural variant
Copy-number change of dosage-sensitive gene
Deletion
Unmasking of a recessive allele
Deletion
*Recessive mutation
*
b Genes that overlap a structural variant
Inversion
Disruption of dosage-sensitive gene
Deletion or translocation
Disruption of dosage-sensitive gene
c Genes that flank a structural variant
Position effect alters expression or regulation of dosage-sensitive gene
Regulatory element
Deletion
Unmasking of a functional polymorphism
Deletion
Functional polymorphism
*
*
d Genes that are involved in complex disorders
No phenotype
Combination of susceptibility variants can cause complex genomic disorders
Inversion
Deletion
Complex disease phenotype
No phenotype

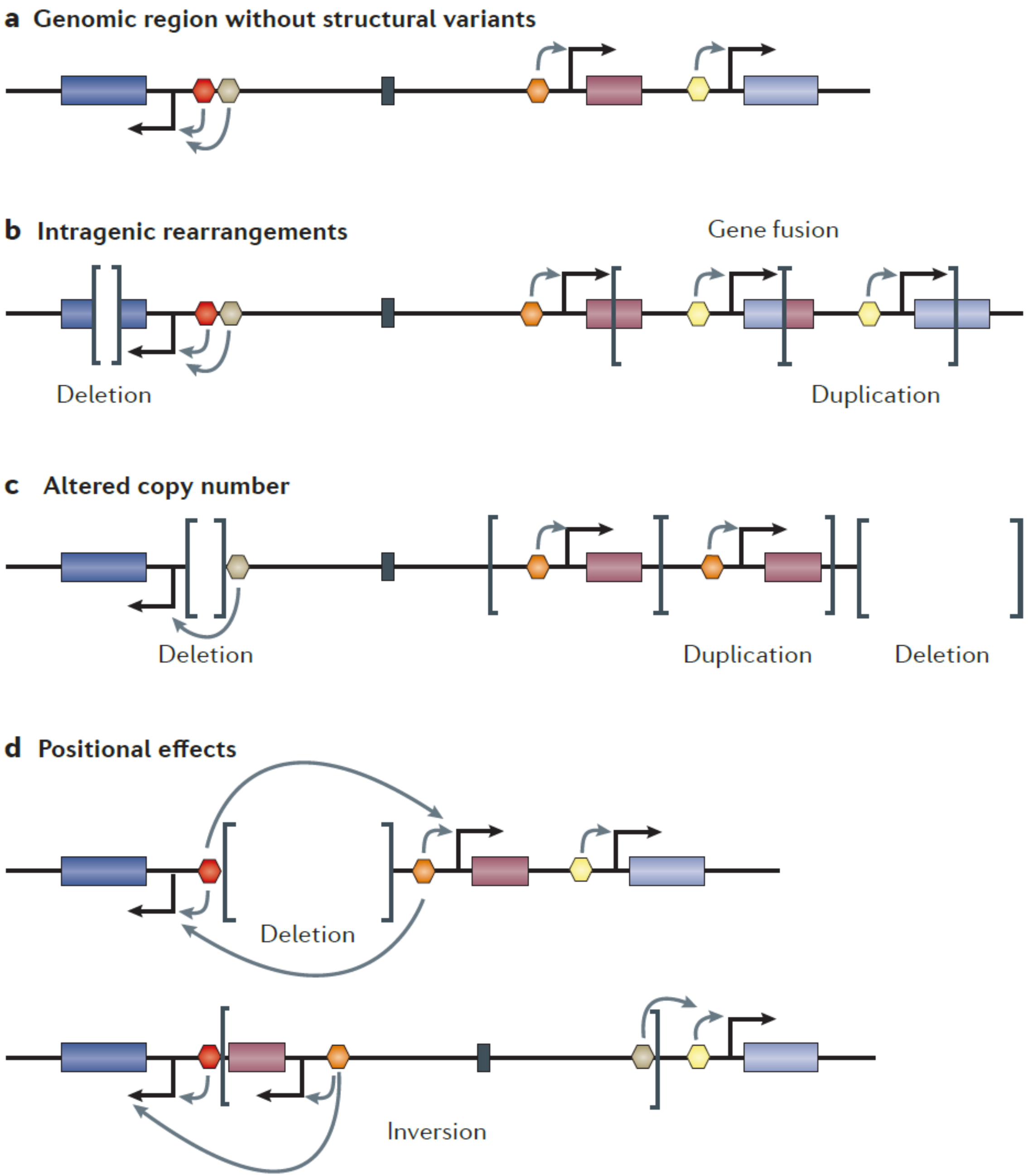
Influence of structural variants on phenotype. Structural variants can be benign, can have subtle influences on phenotypes (for example, they can modify drug response), can predispose to or cause disease in the current generation (for example, owing to inversion, translocation or microdeletion that involves a disease-associated gene), or can predispose to disease in the next generation.

a| Dosage-sensitive genes that are encompassed by a structural variant can cause disease through a duplication or deletion event (upper panel; a deletion is shown here). Dosage-insensitive genes can also cause disease if a deletion of the gene unmasks a recessive mutation on the homologous chromosome (lower panel).

b| Genes that overlap structural variants can be disrupted directly by inversion (upper panel), translocation or deletion (lower panel), or copy-number variant breakpoints (not shown), which leads to the reduced expression of dosage-sensitive genes. Breakpoints that disrupt gene structures can also lead to the formation of new transcripts through gene fusion or exon shuffling (not shown).

c| Structural variants that are located at a distance from dosage-sensitive genes can affect expression through position effects. An example is shown in the upper panel. A deletion of important regulatory elements can alter gene expression; similar effects could result from inversion or translocation of such elements. Alternatively, deletion of a functional element could unmask a functional polymorphism within an effector (lower panel), which could have consequences for gene function.

d| Structural variants can function as susceptibility alleles, where a combination of several genetic factors are required to produce the phenotype. This is illustrated by an example of two structural variants that, individually, do not produce a phenotype. However, in combination they contribute to a complex disease state.



Functional consequences of structural variants.

a| **Genes (boxes)** are regulated by the collective and combinatorial input of regulatory elements, including ***tissue-specific enhancers (hexagons, with different colours indicating tissue-specificity, and arrows pointing to the target gene)*** and ***insulators (black rectangles), which block the activity of regulatory elements***.

b–d| Structural variants (shown by square brackets) can have phenotypic consequences by altering coding regions. For example, they can remove part of a coding region or fuse different coding regions after a duplication, resulting in aberrant transcripts (b). Alternatively, deletions or duplications can lead to altered doses of otherwise functionally intact elements (c), resulting in altered regulatory input (left) or altered gene copy number (right). Structural variants can also affect the expression of genes outside of the variants (that is, a positional effect) (d), thus resulting in a gain or loss of regulatory inputs.

Extensive variation in gene content - CNV

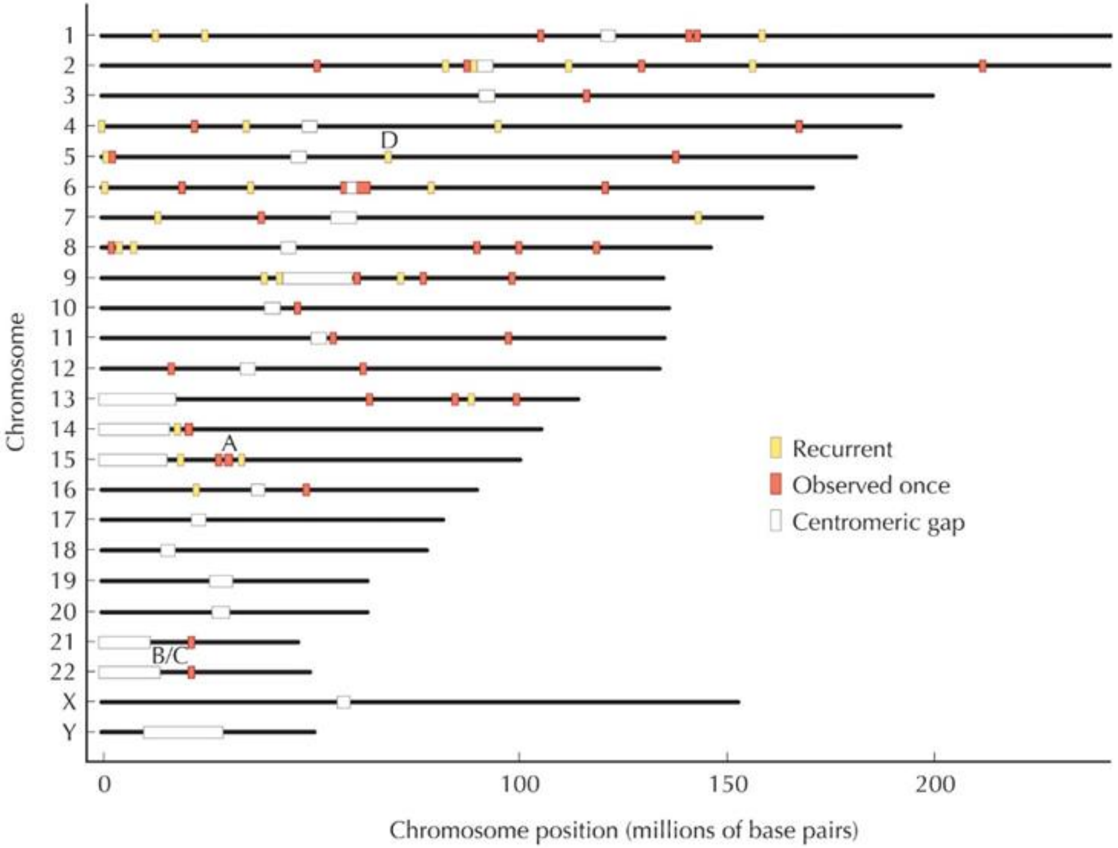
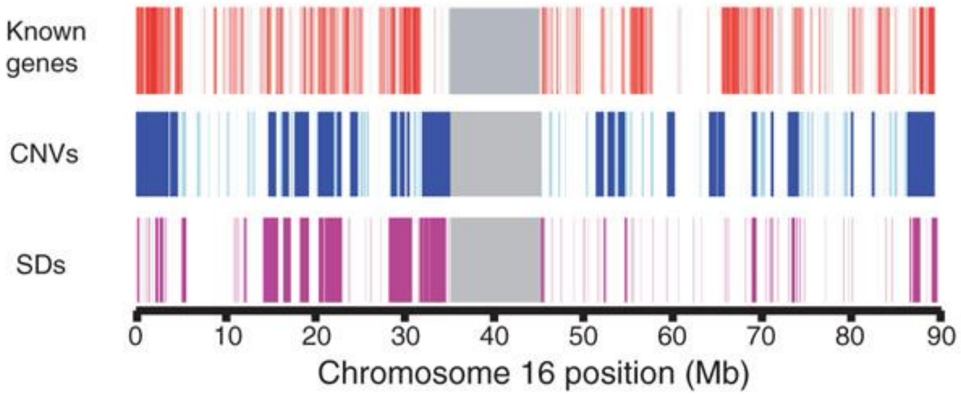


FIGURE 13.32. There is extensive variation in gene content within the human population. DNA from 20 individuals revealed genome-wide variation in copy number at 76 loci, which reflects the presence of deletions and duplications. These polymorphisms each average 465 kb in length and contain 70 known genes in total. Some polymorphisms coincide with loci that have a high rate of chromosomal rearrangement, causing inherited diseases (A, Prader–Willi and Angelman syndromes; B, cat eye syndrome; C, DiGeorge/velocardiofacial syndrome; D, spinal muscular atrophy). This survey should detect most large-scale deletions and insertions but will miss smaller rearrangements.

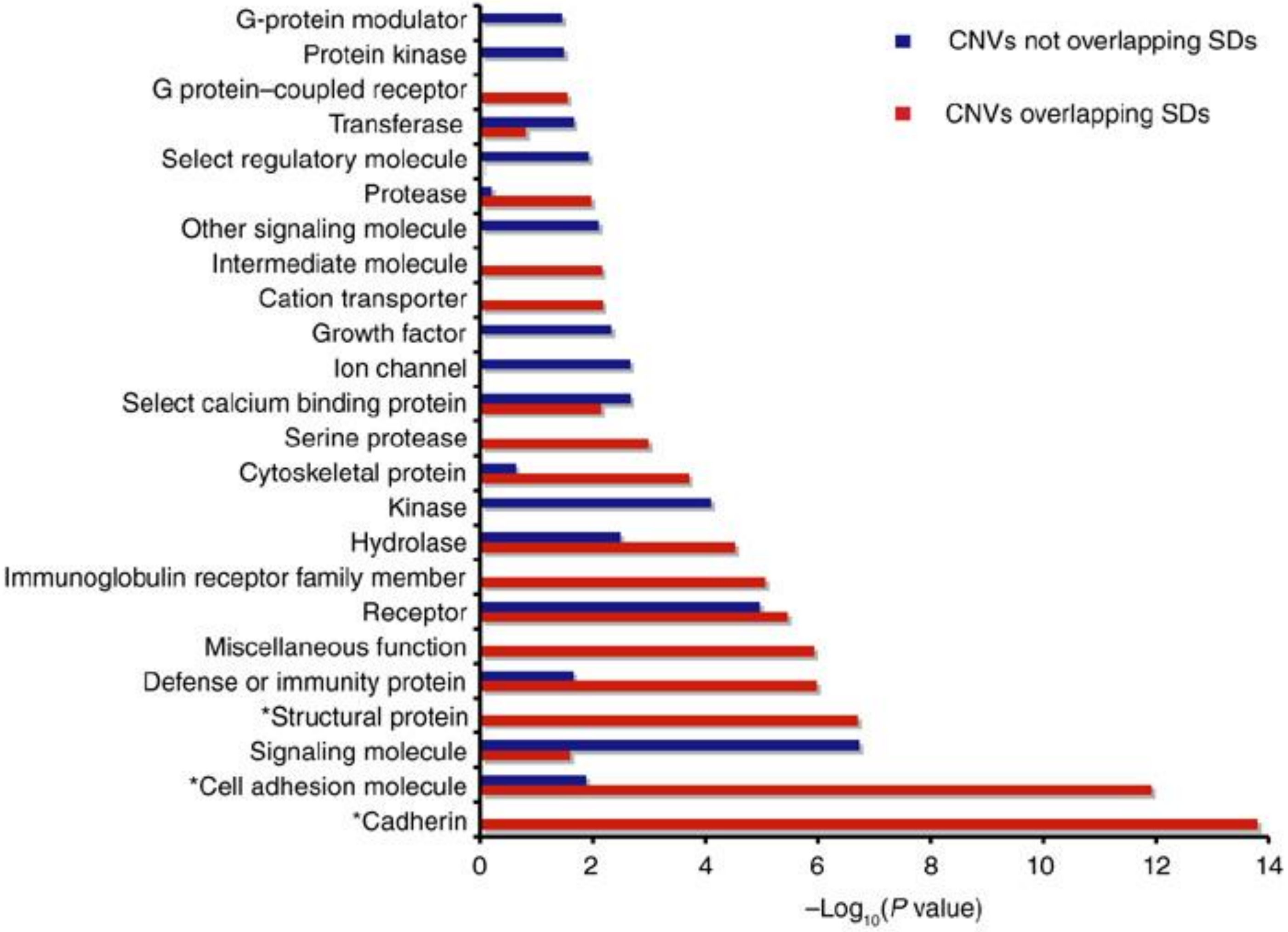
13.32, modified from Sebat J. et al., *Science* **305:** 525–528, © 2004 American Association for the Advancement of Science

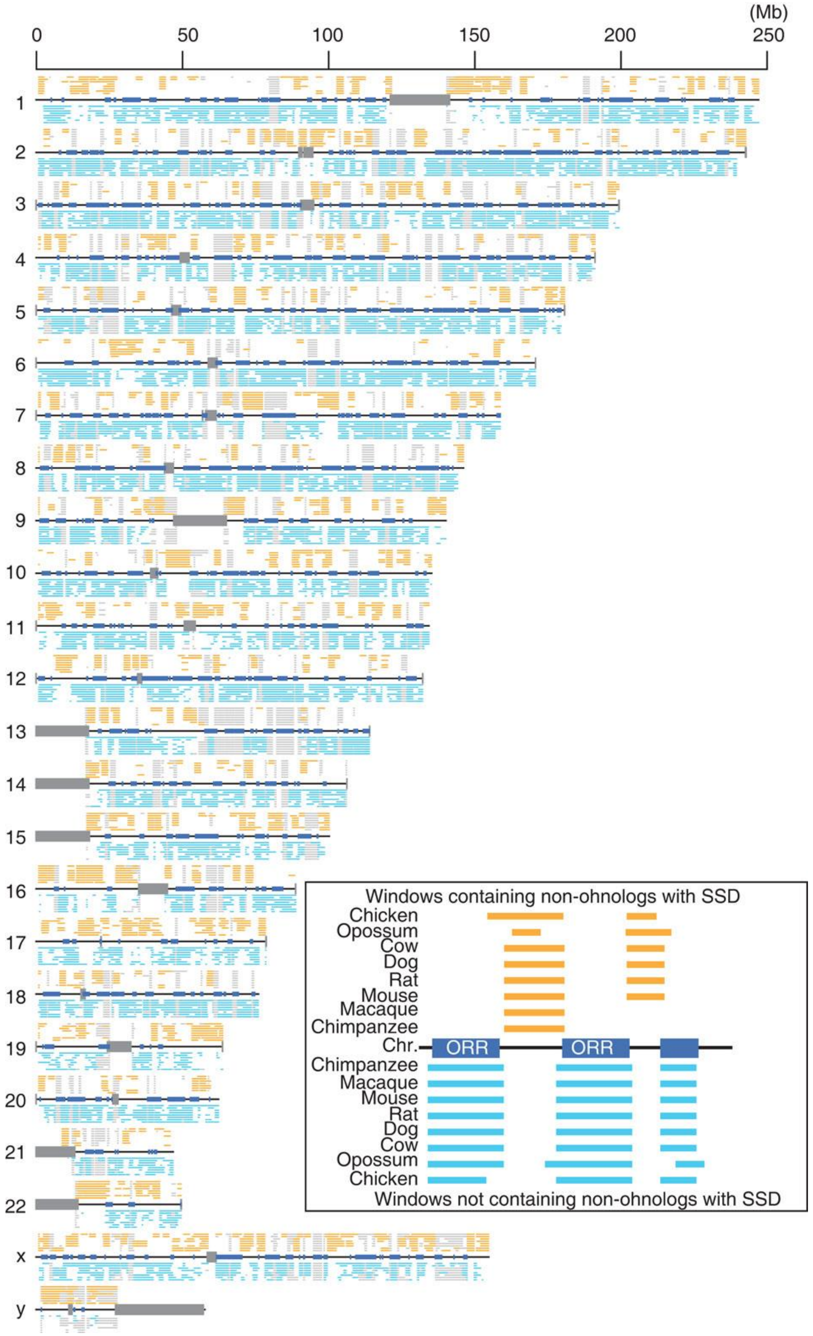
Evolution © 2007 Cold Spring Harbor Laboratory Press

Copy-number variation along human chromosome 16



Functional annotation of copy-number variation gene content





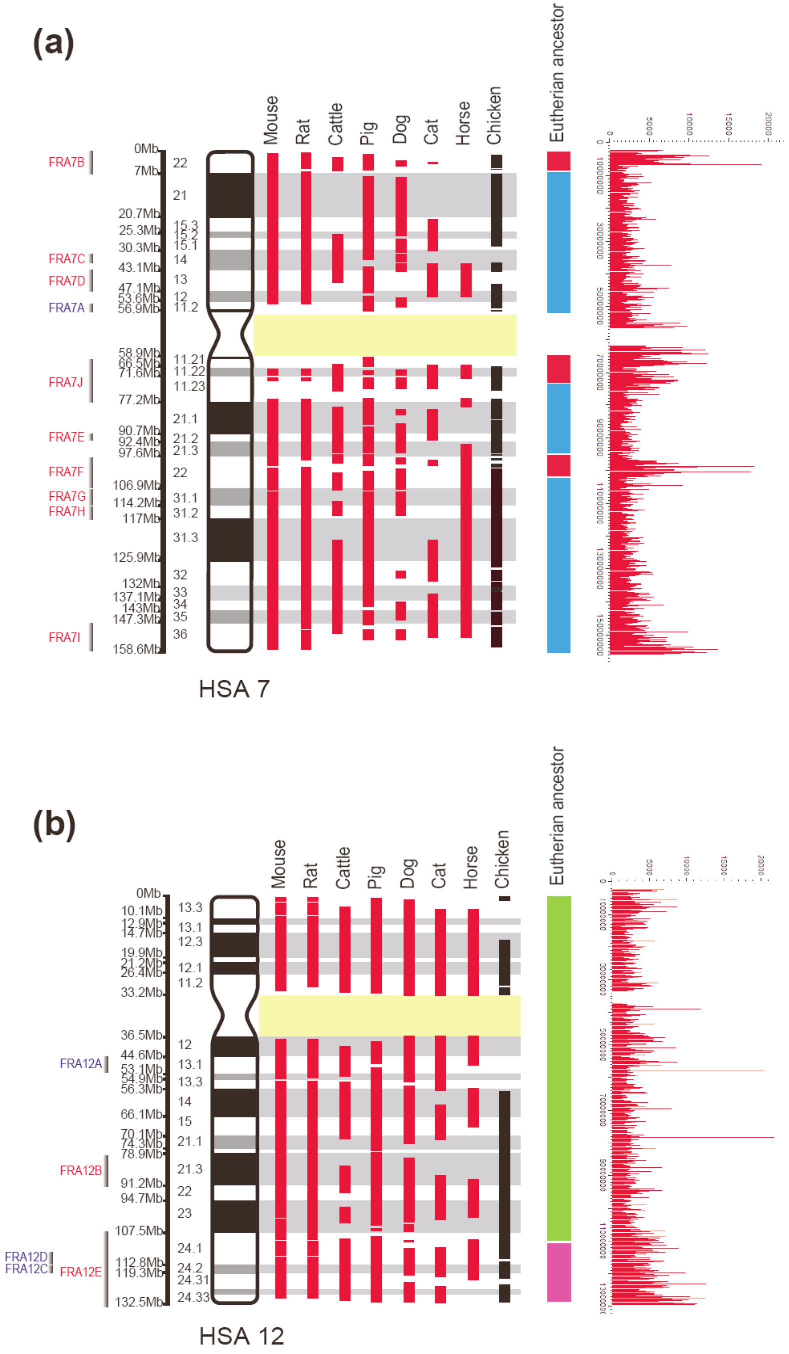
Small-scale gene duplication deserts in humans.

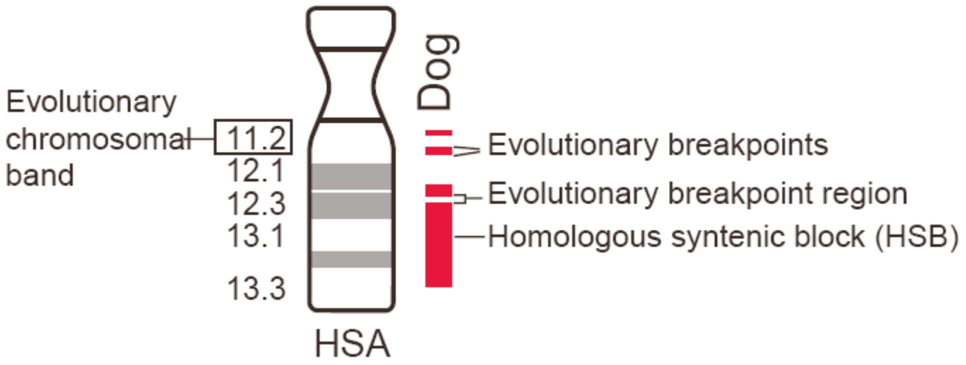
Approximately ***25% of the human genome*** consists of ***gene-poor regions greater than 500 kb***, termed **gene deserts.**

Mammalian chromosomal evolution is driven by regions of genome fragility

-there is a striking correspondence between fragile site (FRA) location, the positions of evolutionary breakpoints, and the distribution of tandem repeats throughout the human genome, which similarly reflect a non-uniform pattern of occurrence,

-certain chromosomal regions in the human genome have been repeatedly used in the evolutionary process. As a consequence, the **genome is a composite of fragile regions prone to reorganization that have been conserved in different lineages, and genomic tracts that do not exhibit the same levels of evolutionary plasticity**.





Evolutionary chromosomal band
11.2
12.1
12.3
13.1
13.3
HSA
Dog
Evolutionary breakpoints
Evolutionary breakpoint region
Homologous syntenic block (HSB)

Kromosomska sintenija (Synteny)

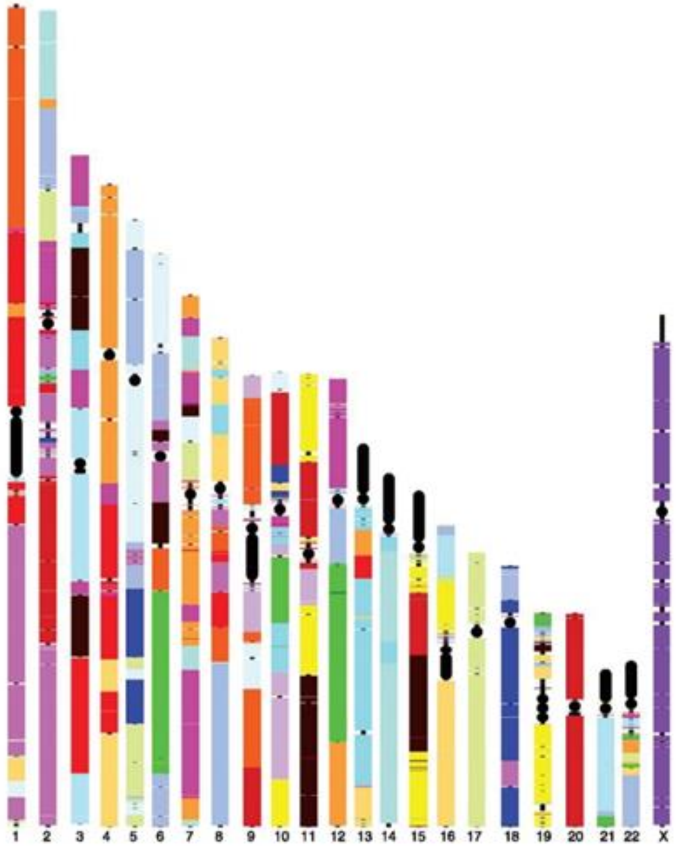


FIGURE 13.33. Gene order is conserved across wide evolutionary distances. The colored segments show blocks of genome that have maintained the same order between mouse and humans. Each color corresponds to a mouse chromosome, overlaid onto the human chromosomes. Note that gene content on the X chromosome is completely conserved (far right).

13.33, reprinted from Eichler E.E. et al., *Science* **301:** 793–797, © 2003 American Association for the Advancement of Science

Evolution © 2007 Cold Spring Harbor Laboratory Press

Synteny = The term synteny was originally defined to mean that two gene loci share the same chromosome. In a genomic context we refer to syntenic regions if both sequence and gene order is conserved between two (closely related) species.

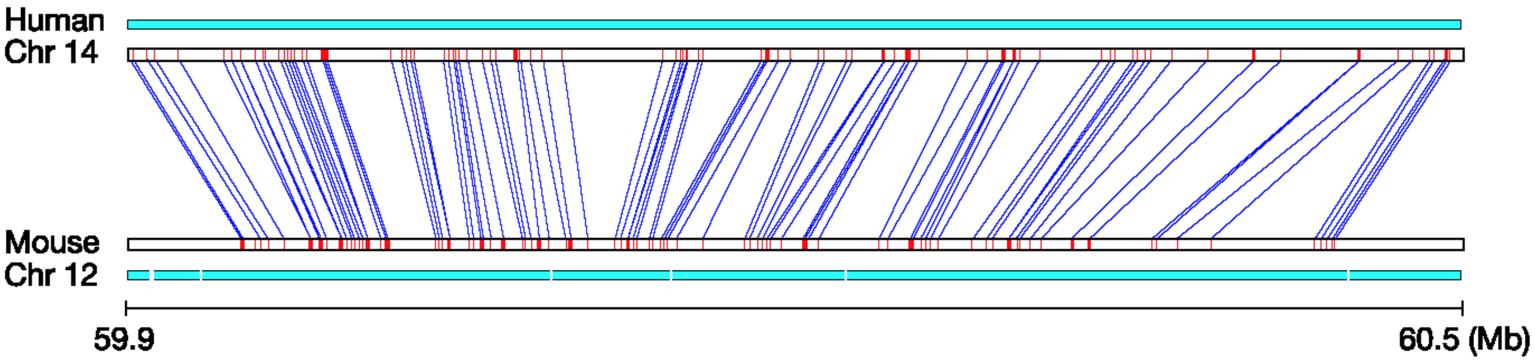
linked genes = genes that reside on the same chromosome.

conserved synteny = a group of linked genes that are highly conserved and hypothesized to be homologous.

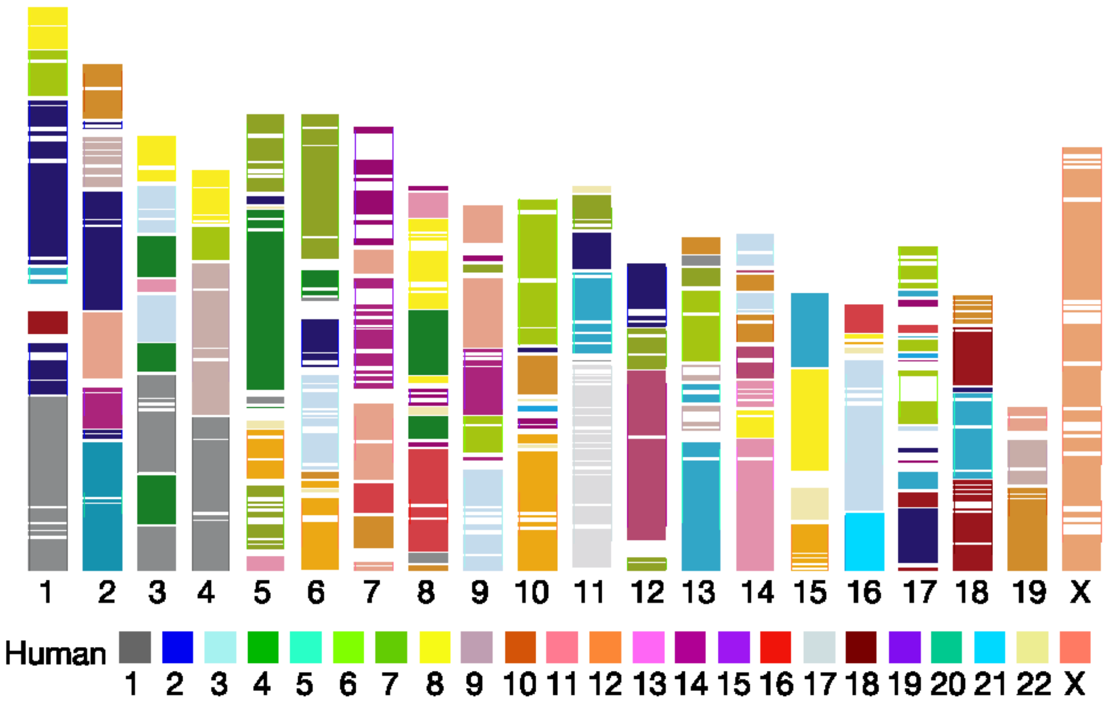
syntenic segment = A group of landmarks that appear in the same order on a single chromosome in each of the two species.

syntenic block = A set of adjacent syntenic segments.

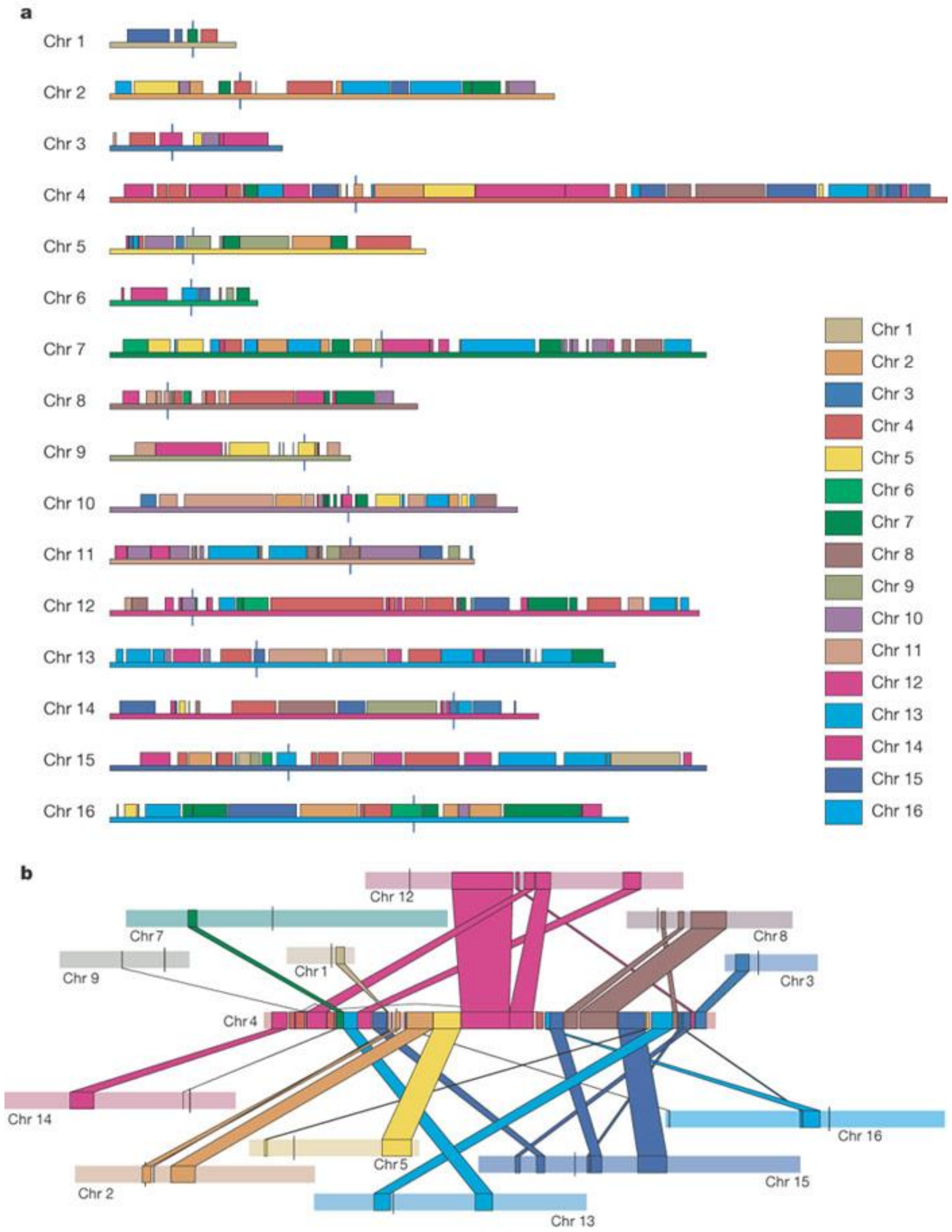
Conservation of synteny between human and mouse.



Segments and blocks >300 kb in size with **conserved synteny in human are superimposed on the mouse genome.**



Whole genome duplication (WGD)

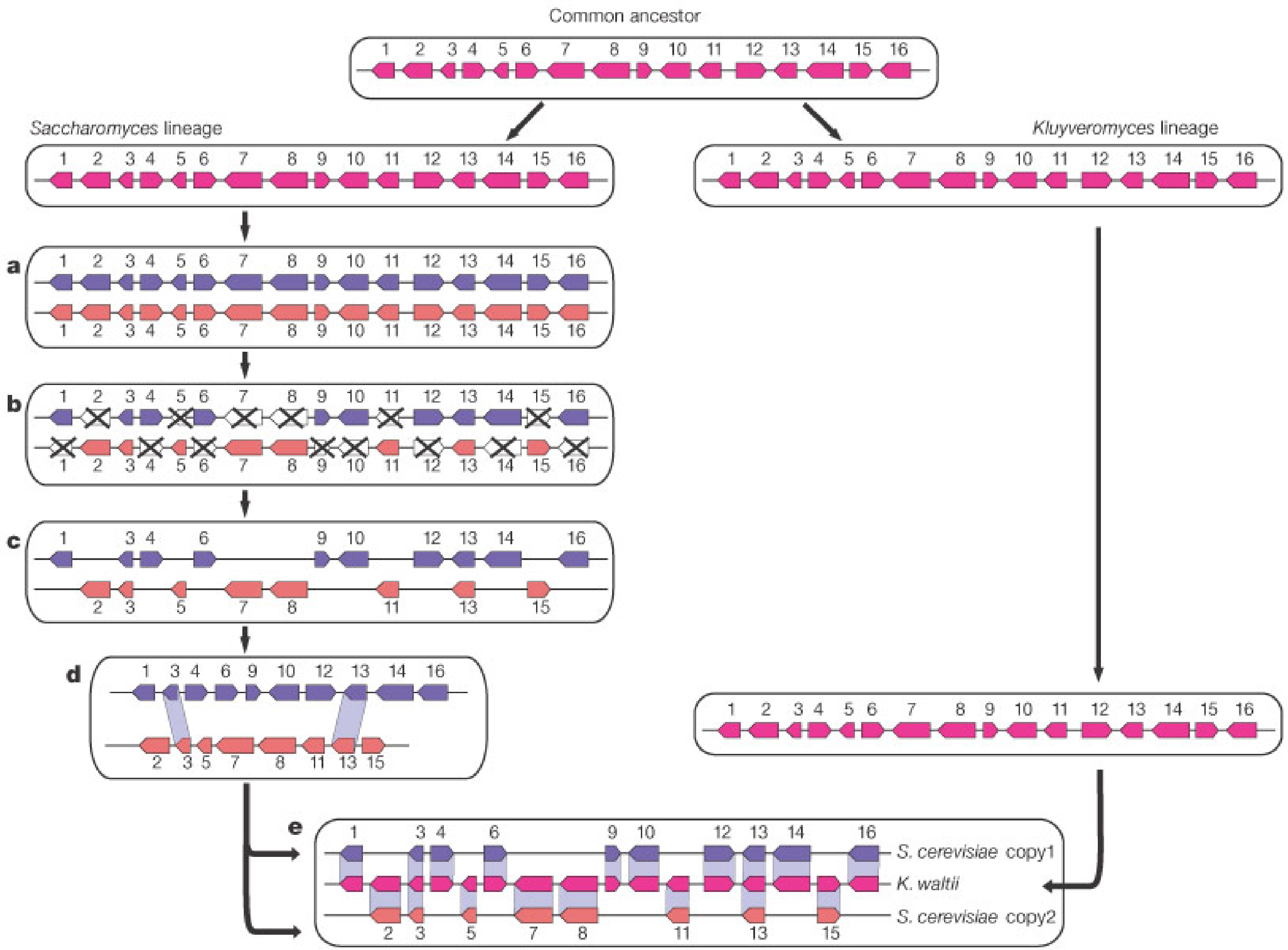


Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae

Duplicated blocks in *S. cerevisiae*.

a, The duplicate mapping in sister regions extends to tile each S. cerevisiae chromosome, revealing complete duplication. Blue vertical bars denote centromeres.

b, Detailed mapping of chromosome 4 with sister regions in other chromosomes.



Model of WGD followed by massive gene loss predicts gene interleaving in sister regions.

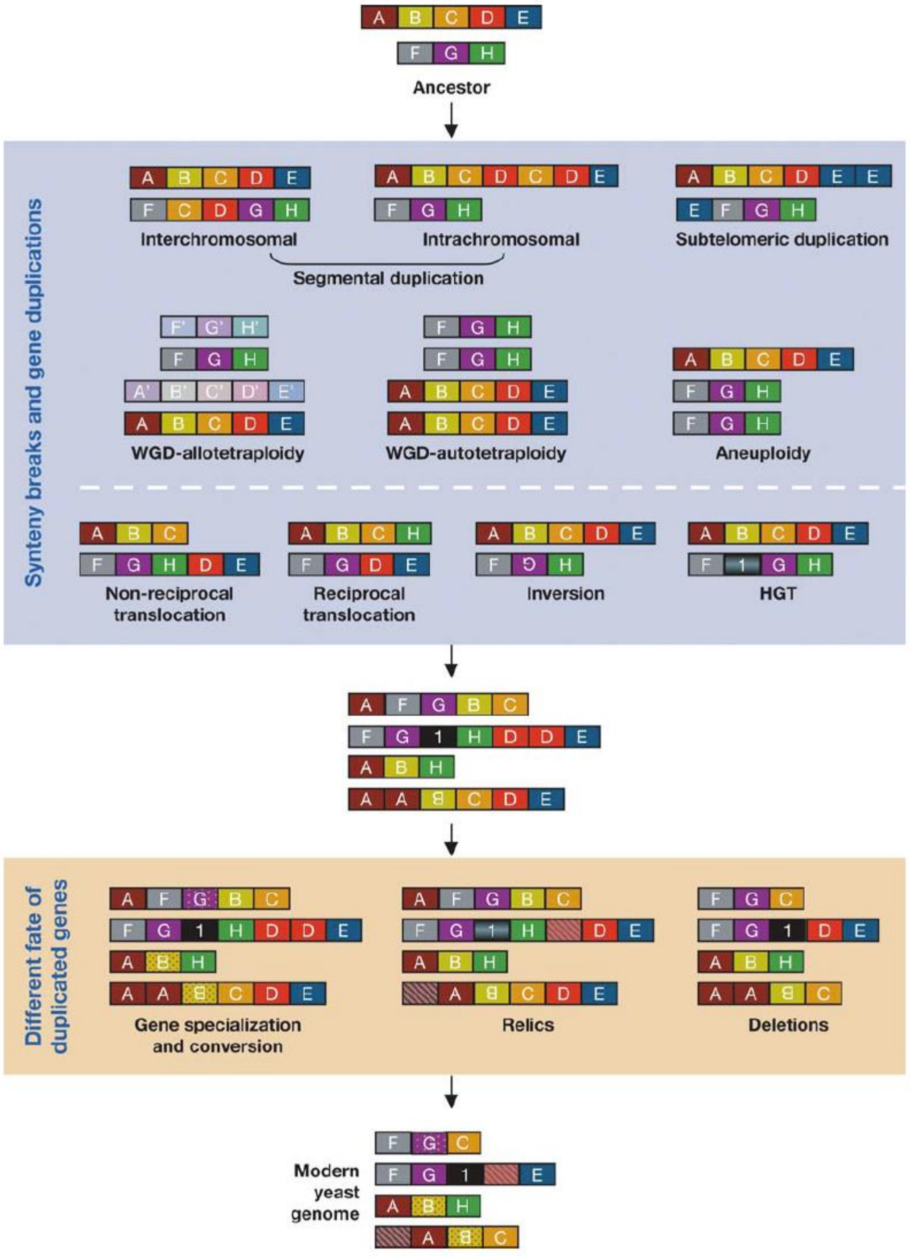
a, After divergence from K. waltii, the *Saccharomyces lineage underwent a genome duplication event, creating two copies of every gene and chromosome.*

b, **The vast majority of duplicated genes underwent mutation and gene loss**.

c, Sister segments retained different subsets of the original gene set, keeping two copies for only a small minority of duplicated genes, which were retained for functional purposes.

d, Within S. cerevisiae, the only evidence comes from the conserved order of duplicated genes (numbered 3 and 13) across different chromosomal segments; the intervening genes are unrelated.

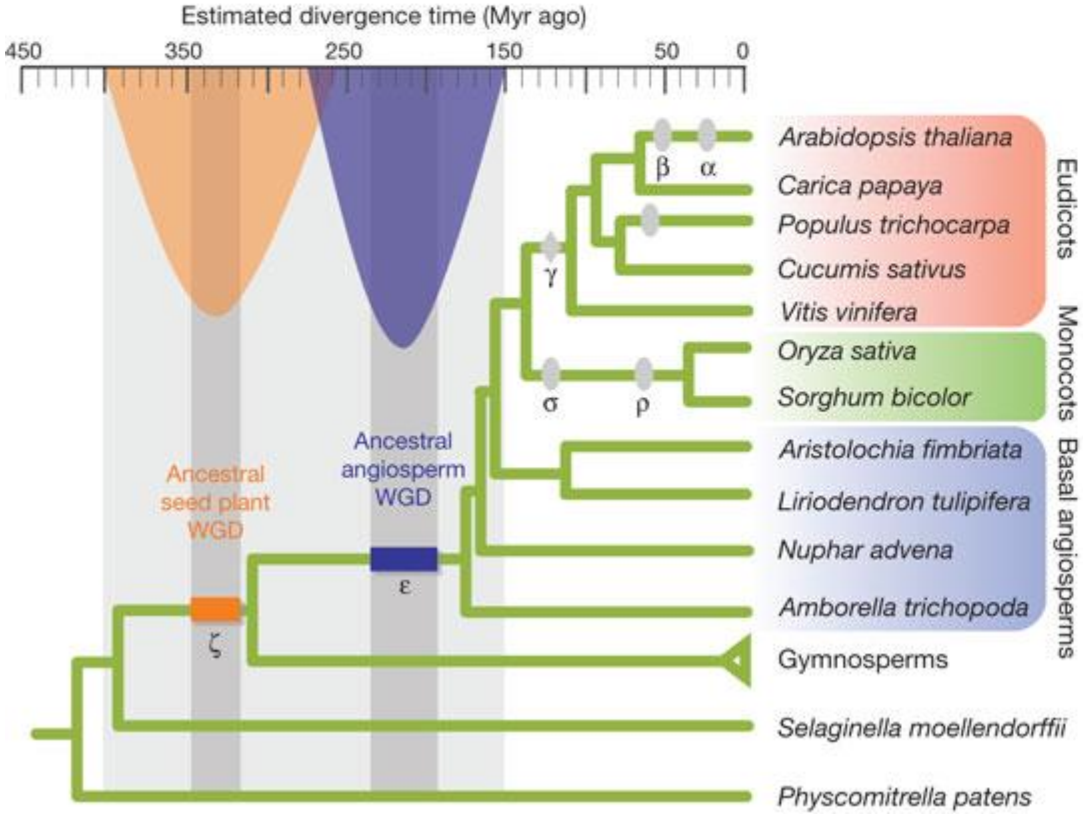
e, Comparison with K. waltii reveals the duplicated nature of the S. cerevisiae genome, interleaving genes from sister segments on the basis of the ancestral gene order.

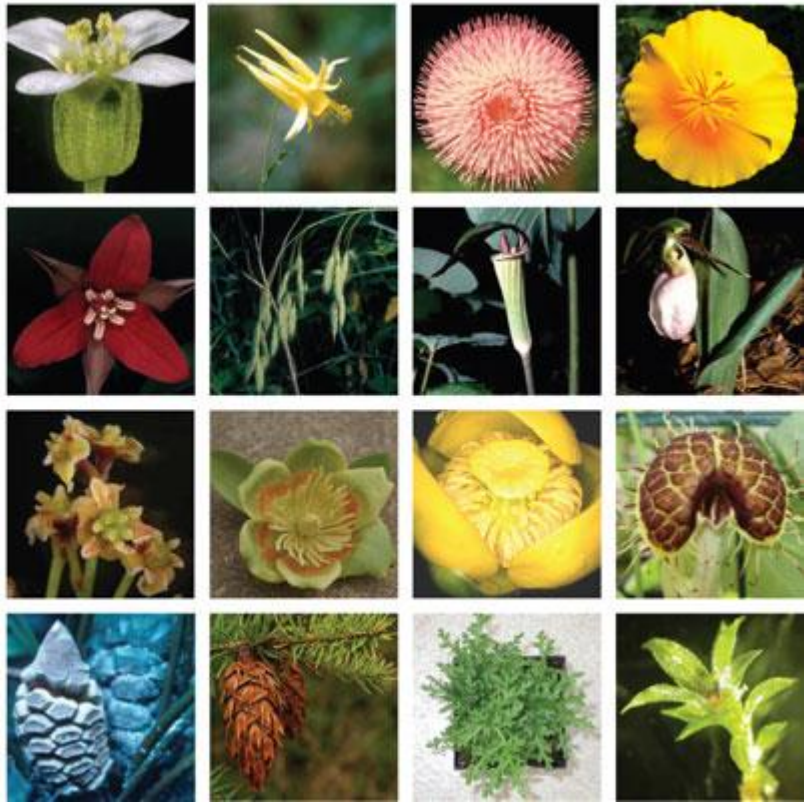


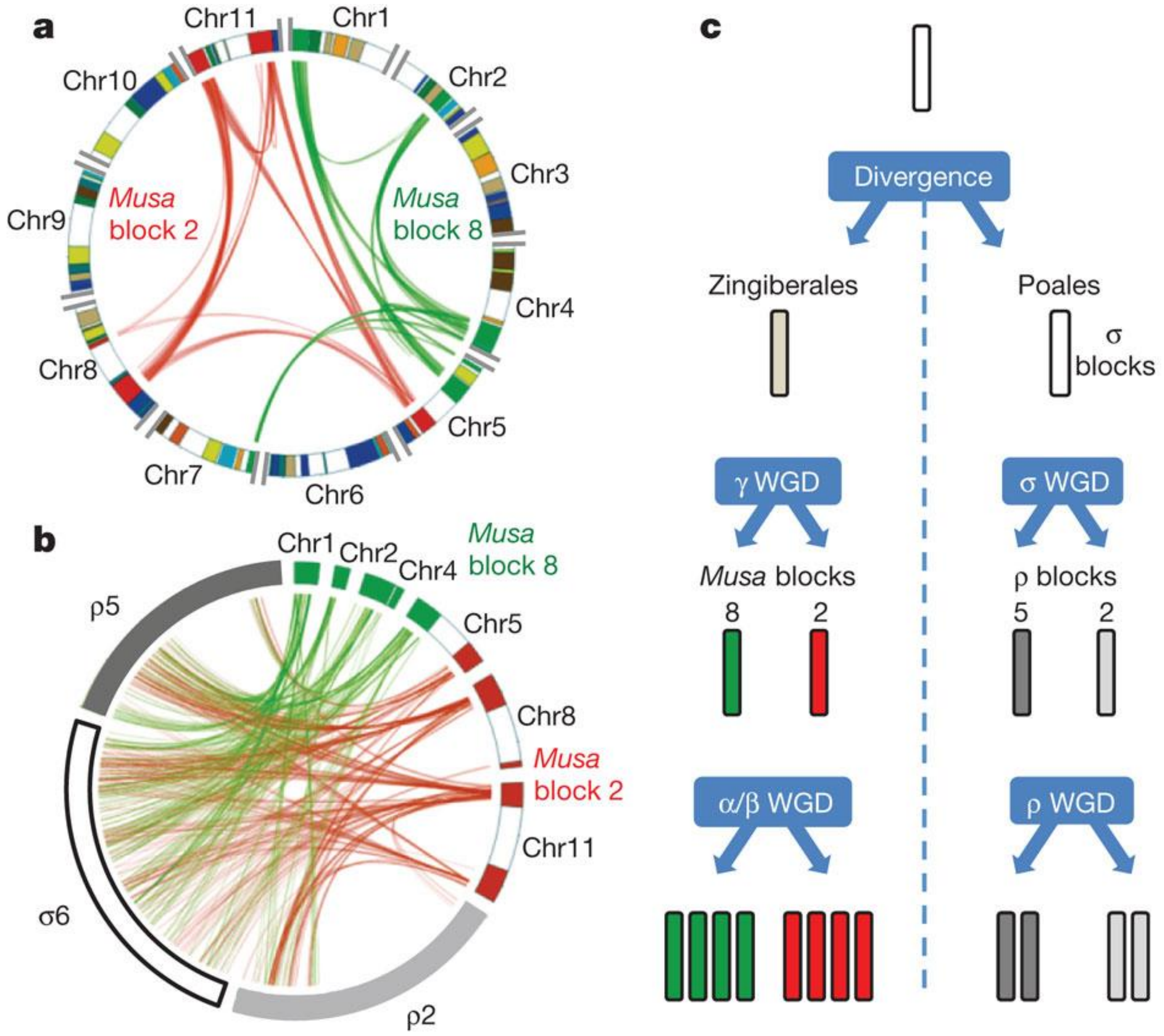
The genomes of yeast are subject to numerous dynamic processes that result in duplications, deletions, rearrangements, mutations, and gene conversions.

Using genome sequences of extant species and their comparisons, we can presume that an ancestral yeast genome can be subject to duplication of the whole genome or segments or both, increasing the copy number of some or all genes. In addition, subtelomeric duplications allow for further expansion of copy number. These duplications can then be subject to a variety of processes including accumulation of mutations, which leads either to the evolution of new functions and gene specialization or to sequence divergence, and inactivation, which leads to relics. Diverged paralogs can result in gene conversion, leading to homogenization of duplicated genes. Massive deletions of small or large segments also occur in these duplications. Aneuploidy is another mechanism of increase or loss of whole chromosomes that can be subject to the same processes. In addition to duplications and their results, **rearrangements of the genome** such as inversions and translocations have been frequent throughout evolutionary history. Although relatively rare in yeast genome evolution, HGT also occurs. All these processes have been and continue to be involved in the **dynamic history of yeast genomes.**

WGD in plants





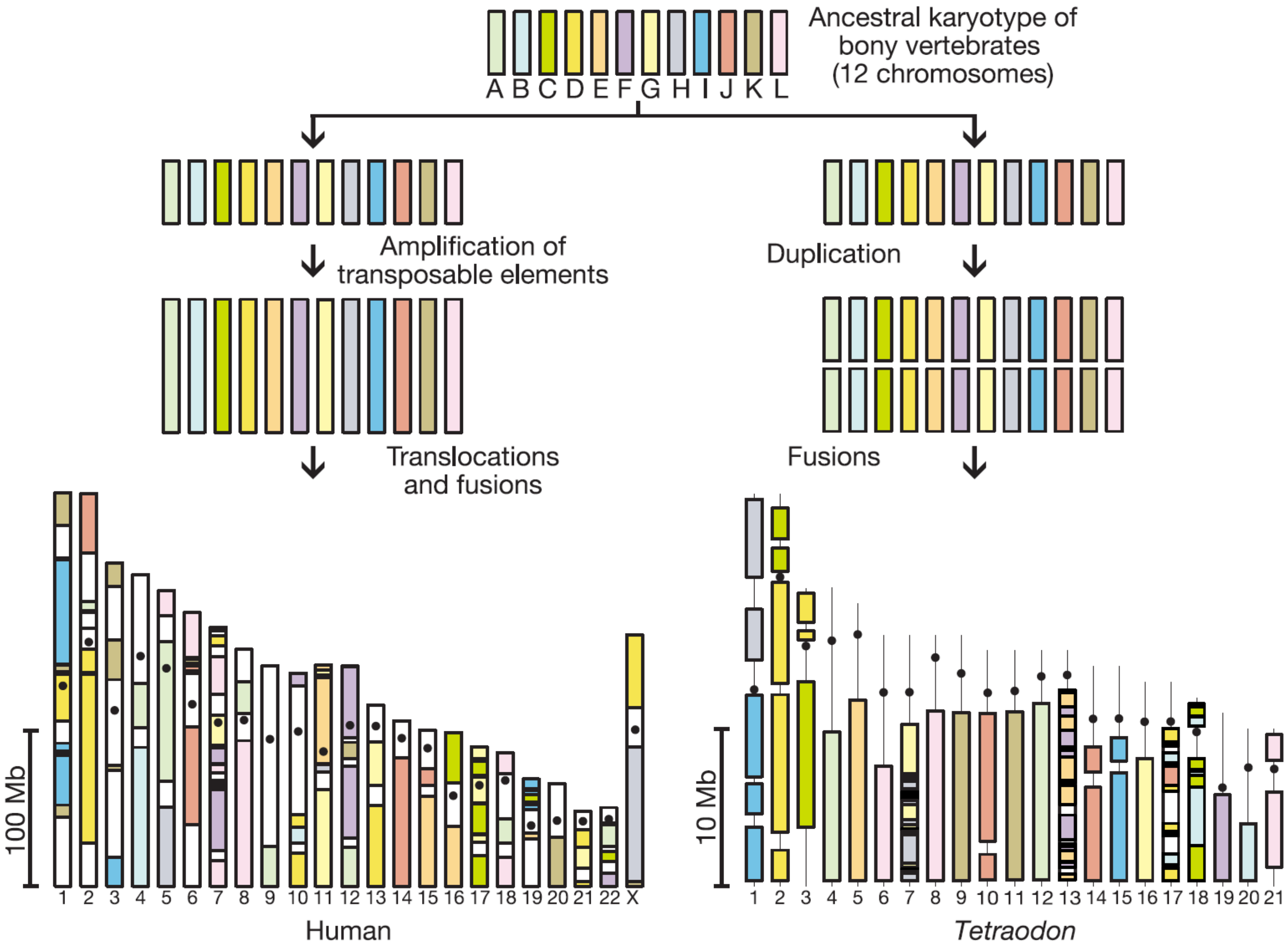


Whole-genome duplication events.

a, ***Paralogous relationships*** between chromosome segments from Musa α/β ancestral blocks 2 (red) and 8 (green). The 12 Musa α/β ancestral blocks are shown in different colours on the circle.

b, ***Orthologous relationship***s of Musa ancestral blocks 2 and 8 with rice ancestral blocks ρ2, ρ5 and σ6. We did not observe a one-to-one relationship between, for instance, Musa α/β ancestral block 2 and one ρ ancestral block, which suggests that the γ and σ duplications are two separate events.

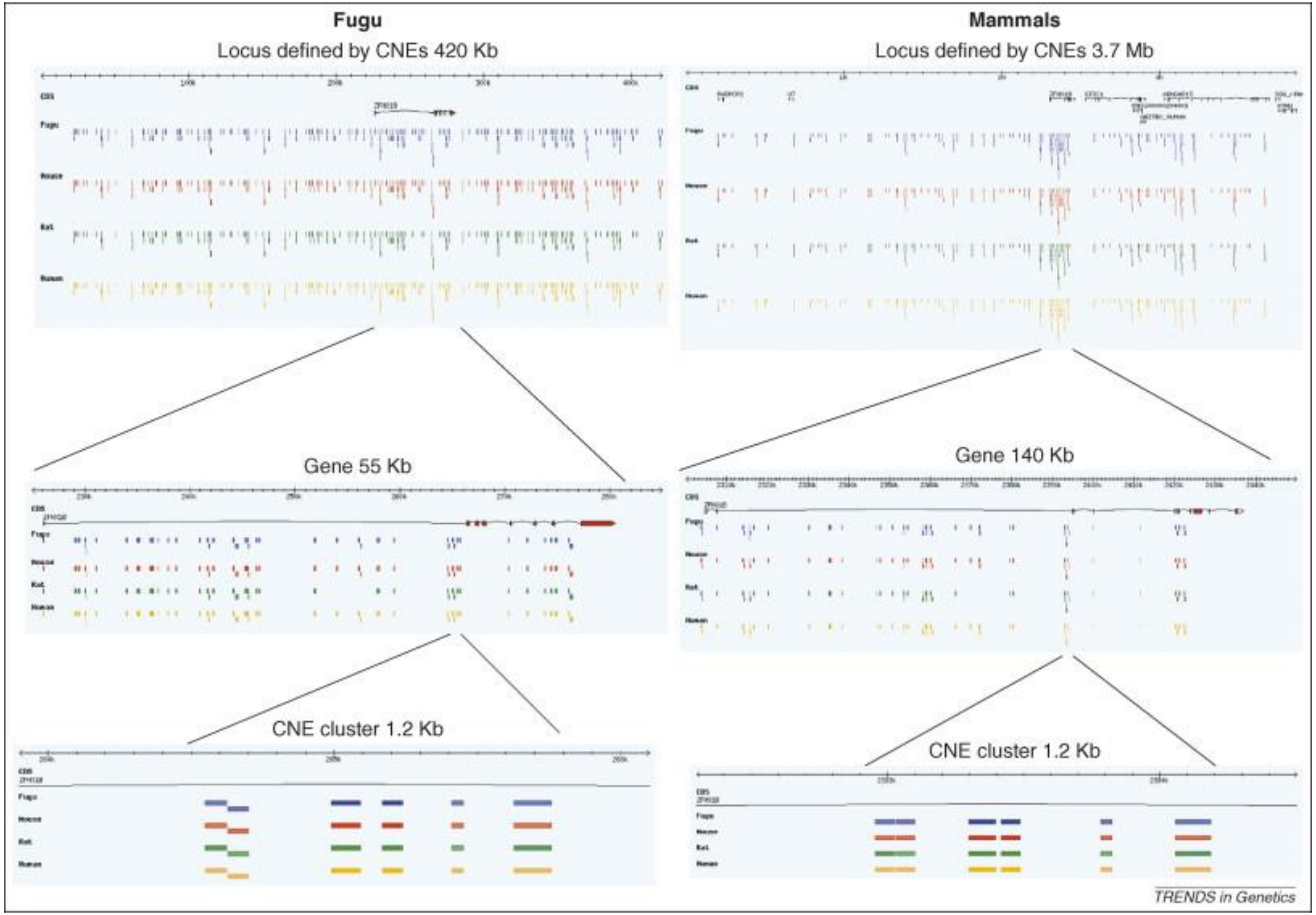
c, Representation of the deduced **WGD event.**


Ancestral karyotype of bony vertebrates (12 chromosomes)
A B C D E F G H I J K L
Amplification of transposable elements
Translocations and fusions
Duplication
Fusions
100 Mb
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Human
10 Mb
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Tetraodon

Proposed model for the distribution of ancestral chromosome segments in the human and the Tetraodon genomes.

The composition of Tetraodon chromosomes is based on their duplication pattern, whereas the composition of human chromosomes is based on the distribution of orthologues of Tetraodon genes. A vertical line in Tetraodon chromosomes denotes regions where sequence has not yet been assigned. With 90 blocks in human compared with 44 in Tetraodon, the complexity of the mosaic of ancestral segments in human chromosomes underlines the higher frequency of rearrangements to which they were submitted during the same evolutionary period.

Ultraconserved elements in the genome: CNEs, UCRs etc.

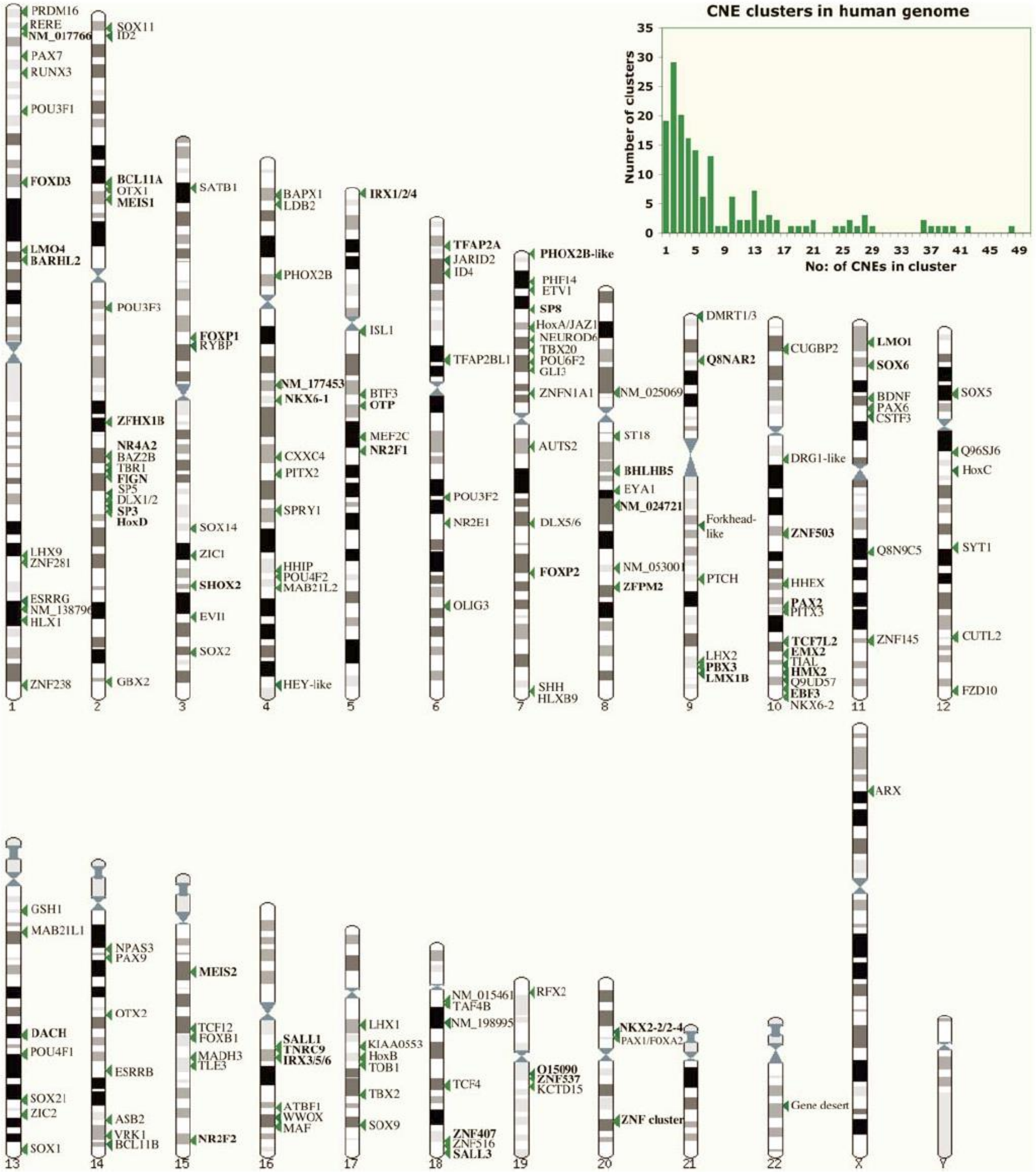


Conserved noncoding elements (CNEs) define large multifunctional regulatory modules.

Diagrammatic representation of the sequence conservation landscape of the ***ZFHX1B locus in the human, mouse, rat and Fugu genomes***. The diagrams on the left show the conservation landscape of the alignments with Fugu as the baseline genome, whereas the diagrams on the right show the equivalent alignments with human as the baseline. Despite a wide distribution of genome sizes, ***the distances between the CNEs across this region in all four species is very similar (1.25 kb ± 50 bp), implying strong spatial constraint even in the absence of sequence conservation.***

Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development

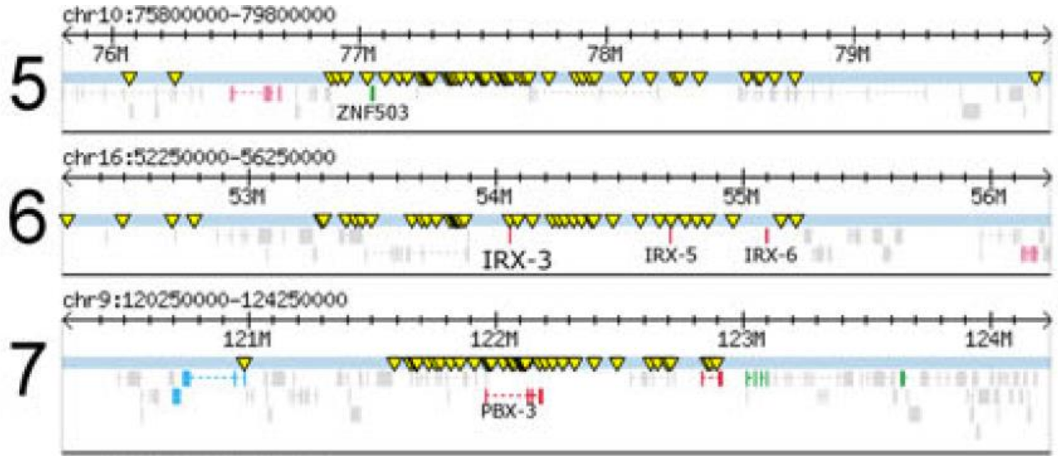
CNE (conserved non-coding element) Clusters Are Found Close to Trans-Dev (regulation of development) Genes in the Human Genome

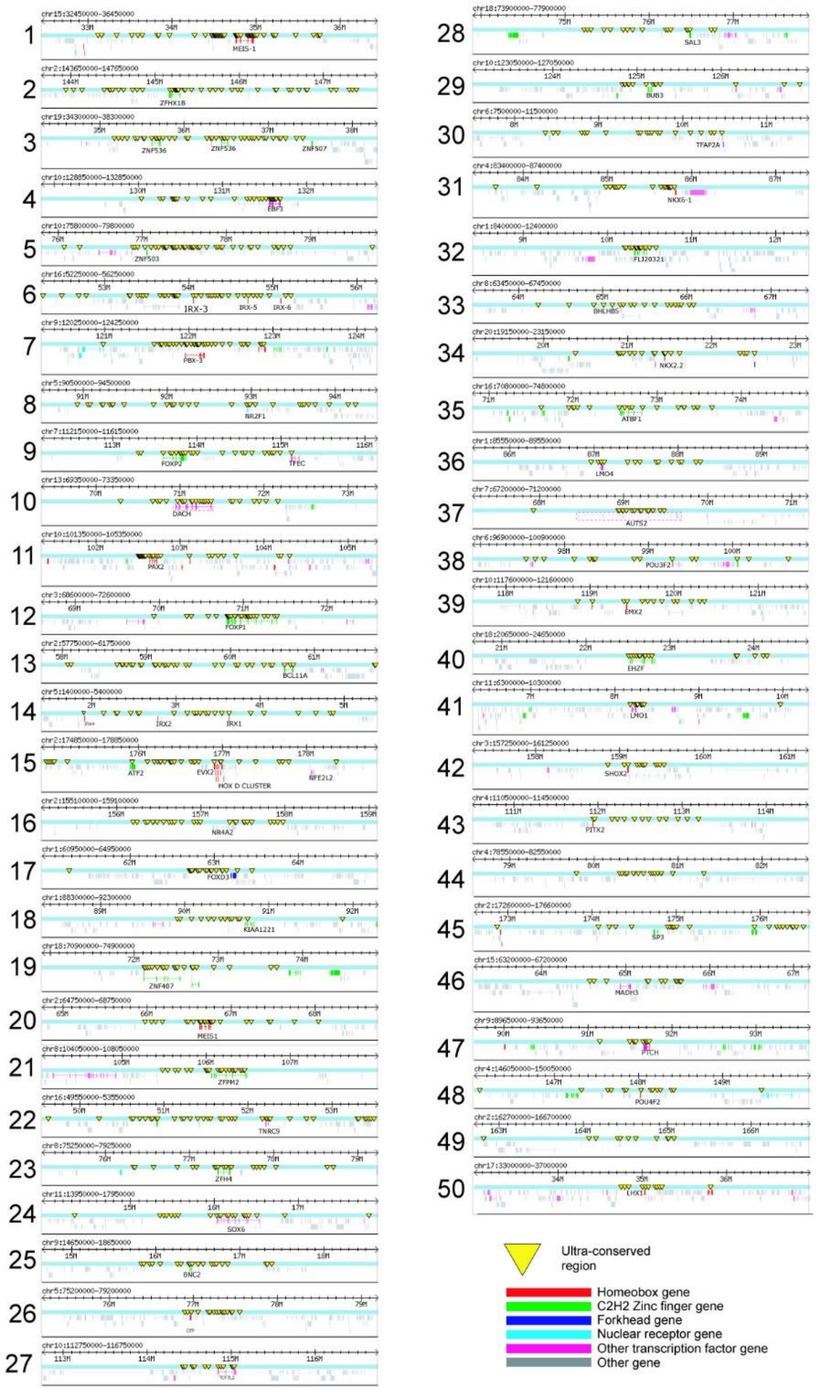


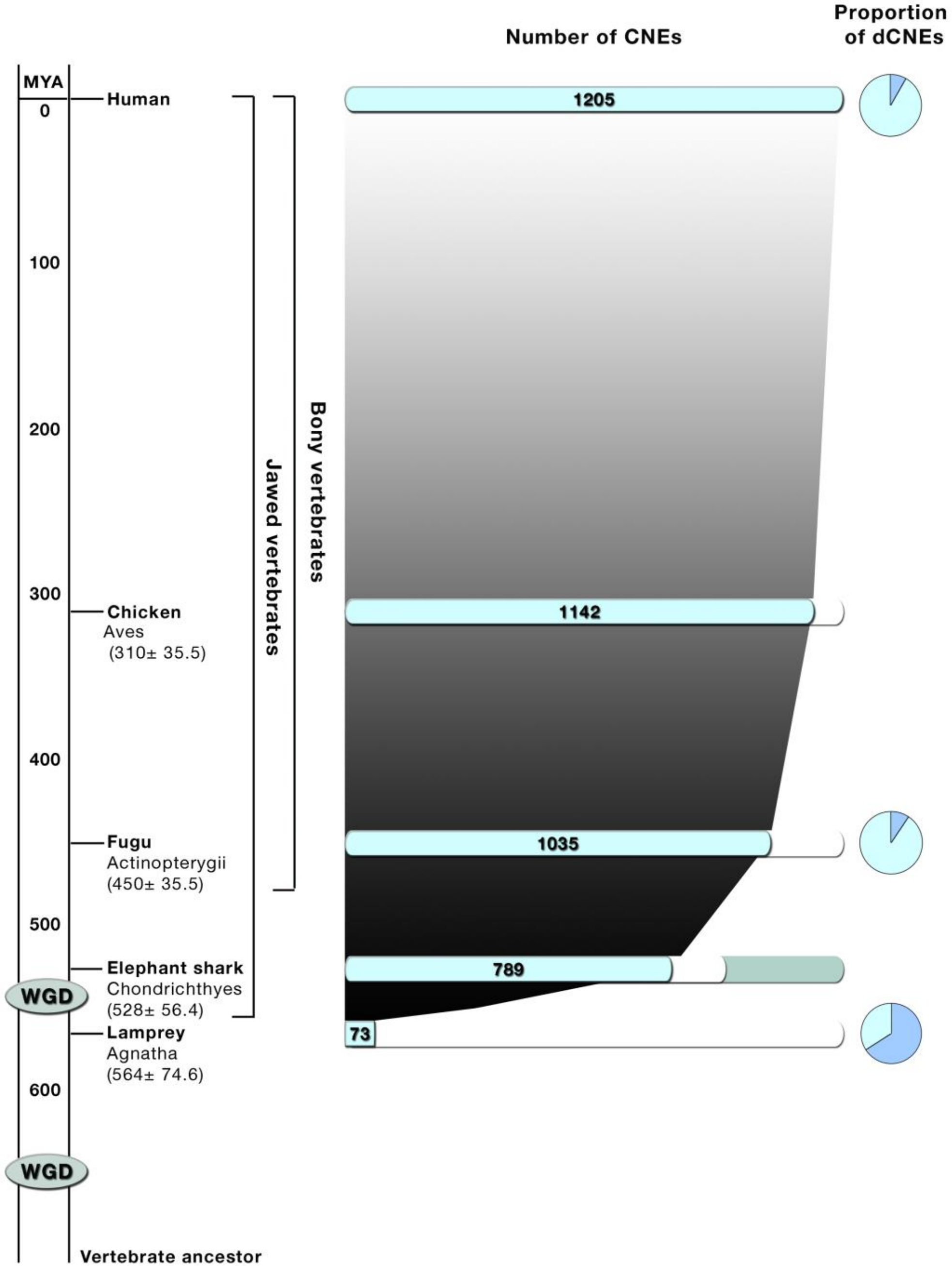
Genomic landscape surrounding the most prominent UCR clusters in the human genome

UCR = ultra-conserved region

UCRs were counted by sliding a 500 kb window along the chromosomes. Overlapping UCR-containing windows were merged into a single cluster span. Each of the regions shows a 4 MB region around the corresponding UCR cluster. The cluster span coordinates correspond to the human genome NCBI build 33 (UCSC hg15, April 2003). **Transcription factor genes** are colored according to structural class. **UCR clusters are visibly correlated with transcription factor genes**; other developmental regulators that do not contain any of the probed protein domains were located manually (boxed), such as the autism susceptibility gene (chromosome 7, number 37) and the DACH gene (chromosome 13, number 10).



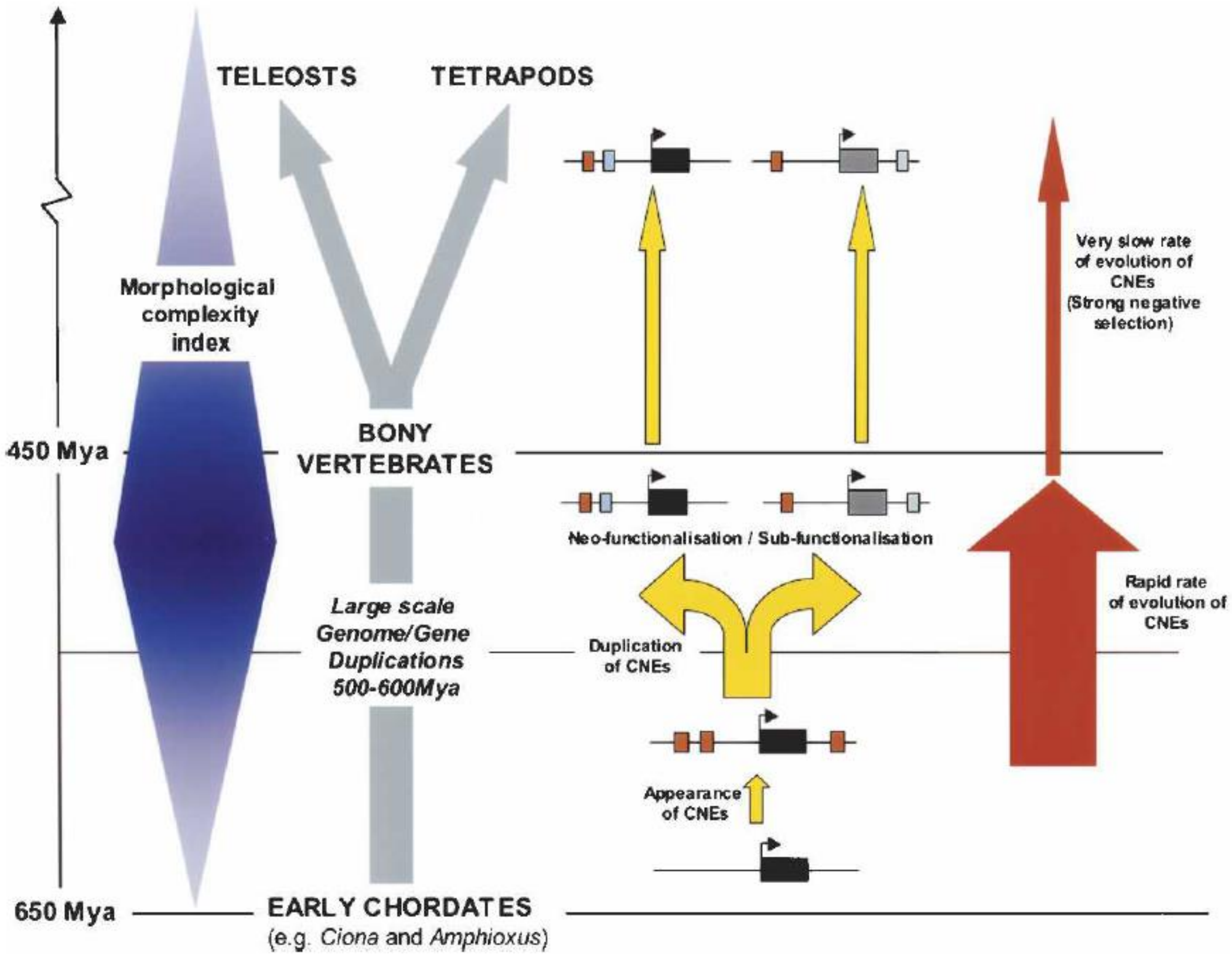




Proposed model of early vertebrate evolution, genome duplications, and expansion of CNE repertoire.

Evolution of CNE repertoire occurred early in the history of vertebrates, coinciding with proposed genome duplication events and the emergence of agnathans. CNEs are barely detectable in invertebrate genomes and therefore must have evolved very early in the ancestral vertebrate, coinciding with whole genome duplication (WGD) events. The lamprey genome possesses a much smaller set of CNEs than sharks and other jawed vertebrates. This suggests that a large number of CNEs evolved and became fixed in all gnathostomes within a relatively short time period, between the emergence of agnathans and Chondrichthyes, coincident with the second proposed large scale genome duplication in jawed vertebrates. The proportion of CNEs that are found to be duplicated in the human genome is shown for each species (proportion of dCNEs); over half of the human CNEs that have matches to lamprey are found to be duplicated in the human genome. Timescale and divergence times taken from. The elephant shark genome has a maximum coverage of 75% and the light blue area represents the unsequenced portion.

CNE=conserved non-coding elements



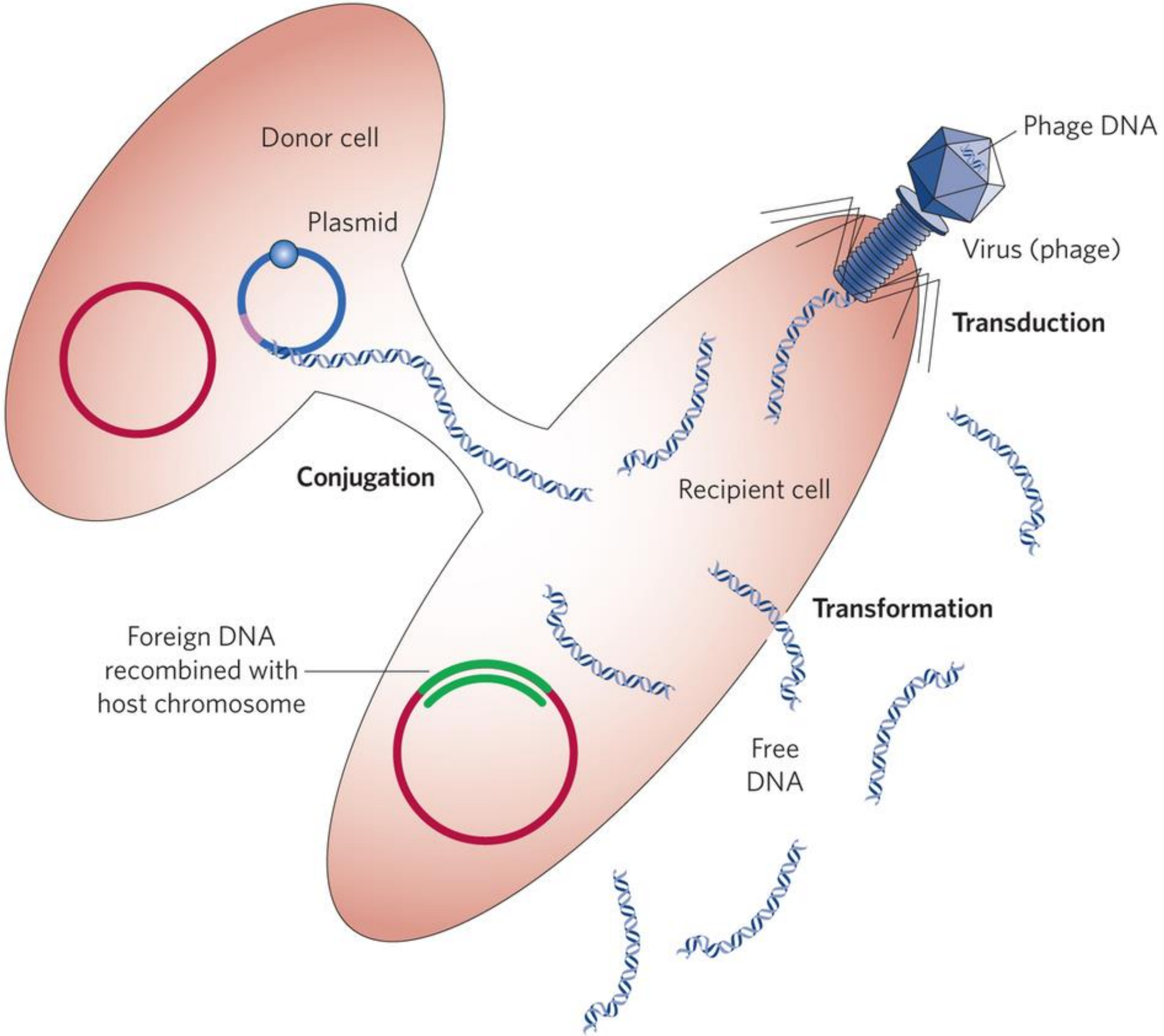
TELEOSTS
TETRAPODS
Morphological complexity index
Very slow rate of evolution of CNEs (Strong negative selection)
450 Mya
BONY VERTEBRATES
Neo-functionalisation / Sub-functionalisation
Rapid rate of evolution of CNEs
Large scale Genome/Gene Duplications 500-600Mya
Duplication of CNEs
Appearance of CNEs
650 Mya
EARLY CHORDATES
(e.g. Ciona and Amphioxus)

Proposed model of CNE evolution in the context of other major genomic events during the early vertebrate radiation.

Modern bony vertebrates evolved from the chordate lineage between 650 and 450 Mya, during a period of rapid morphological change (blue). It is now generally accepted that during this period an early ancestral vertebrate underwent one, or possibly two, whole-genome duplications, generating a greatly increased repertoire of genes, which in turn may have contributed to this increase in morphological complexity. The appearance of CNEs in vertebrate genomes (red boxes adjacent to gene loci, depicted as dark boxes) can be dated prior to these large-scale duplication events, as most of the dCNEs are associated with trans-dev paralogs that derive from these ancient duplications (yellow arrows). The duplication of gene loci together with associated cis-regulatory modules generates the plasticity for genes to develop new functions (neofunctionalization) and/or to perform a subset of the functions of the parent gene (subfunctionalization). This evolution must have occurred rapidly following duplication over a relatively short evolutionary period (~50–150 Myr) during which time dCNEs evolved in length and sequence. In contrast, in the period since the teleost–tetrapod divergence (~450 Mya), dCNEs have had a remarkably slow mutation rate and have remained practically unchanged.

dCNE = duplicated CNE.

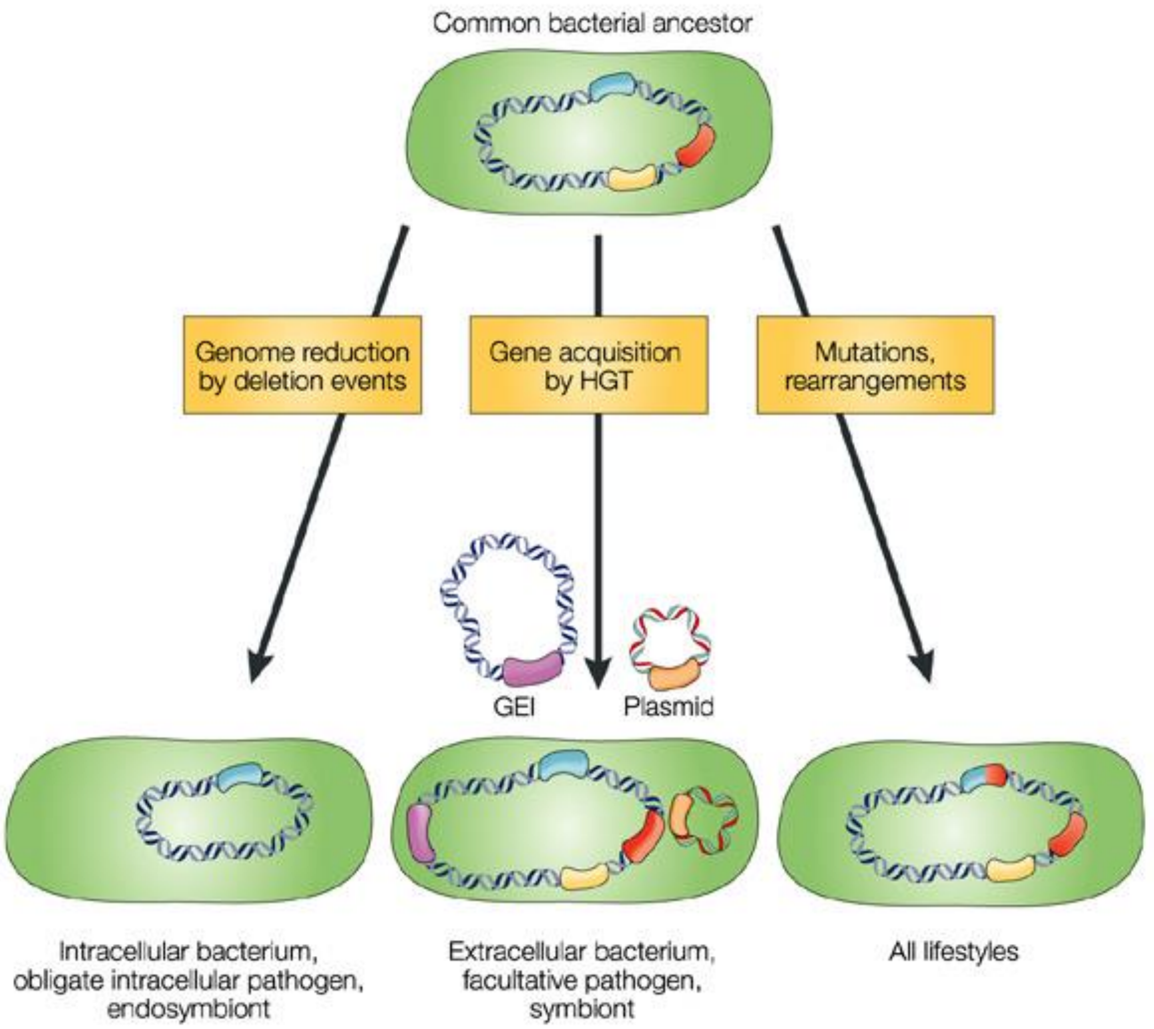
Genome dynamics – gene loss and gene gain (by HGT) in bacteria



Routes of horizontal gene transfer in bacteria and archaea.

Bacteria and archaea can acquire new genes, or variant copies of existing genes, from genetically distant relatives through a process termed **horizontal gene transfer**.

This can occur through the uptake of extracellular DNA (**transformation**), cell-to-cell transfer through surface appendages (**conjugation**) and viral import (**transduction**).


Common bacterial ancestor
Genome reduction by deletion events
Gene acquisition by HGT
Mutations, rearrangements
GEI
Plasmid
Intracellular bacterium, obligate intracellular pathogen, endosymbiont
Extracellular bacterium, facultative pathogen, symbiont
All lifestyles

Nature Reviews | Microbiology

Genomic islands in pathogenic and environmental microorganisms.
Horizontal gene transfer is an important mechanism for the evolution of microbial genomes.

Pathogenicity islands — mobile genetic elements that contribute to rapid changes in virulence potential — are known to have contributed to genome evolution by horizontal gene transfer in many bacterial pathogens.

Increasing evidence indicates that **equivalent elements in non-pathogenic species — genomic islands — are important in the evolution of these bacteria, influencing traits such as antibiotic resistance, symbiosis and fitness, and adaptation in general.**

non-pathogenic species: commensal, symbiotic and environmental bacteria.

Pathogenicity island

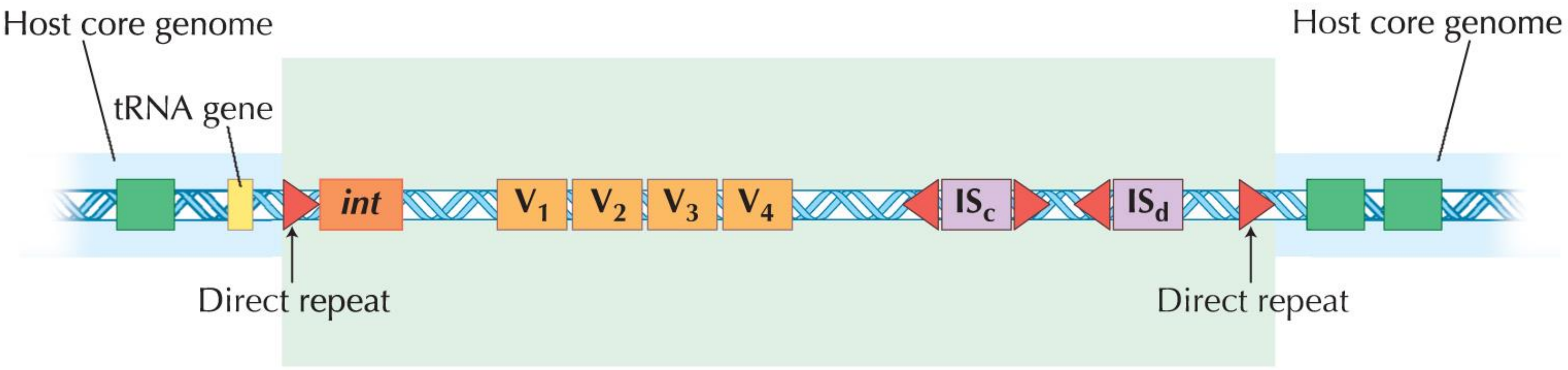


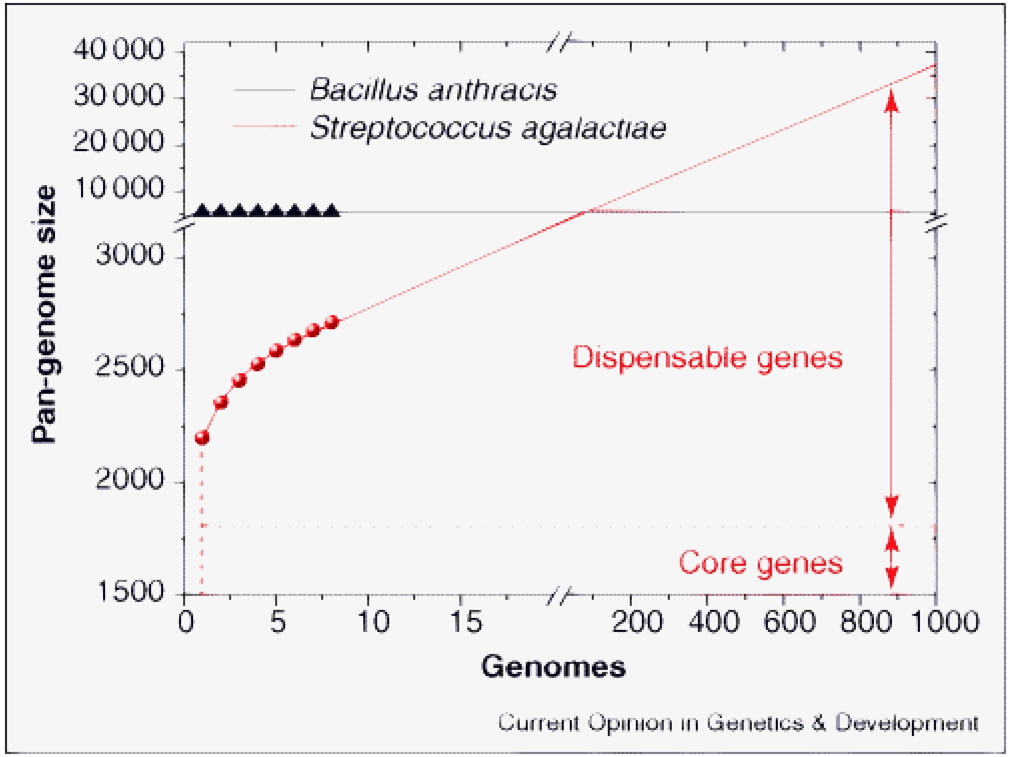
FIGURE 7.21. Diagram of a pathogenicity island (within the *large green shaded region*). Flanking the island are core host genes (*small blue shaded regions*). Immediately flanking the island on the *left* is a tRNA gene—many islands are found near or within tRNA genes. At the edges of the island are direct repeats (*red triangles*). Inside the island are virulence genes (*orange boxes*) and some insertion sequences (*purple boxes*).

7.21, adapted from Schmidt H. et al., *Clin. Microbiol. Rev.* **17:** 14–56, © 2004 American Society for Microbiology

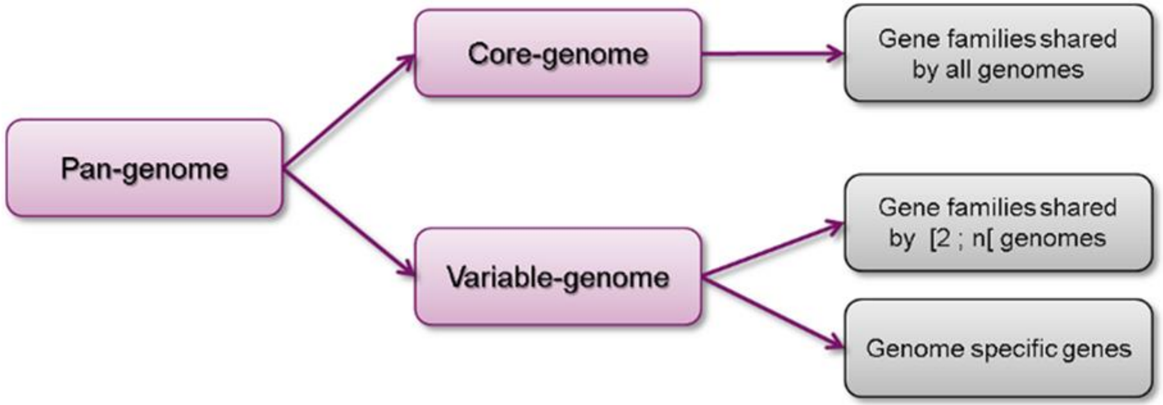
Evolution © 2007 Cold Spring Harbor Laboratory Press

Table 1 | **Examples of genomic islands and their encoded functions**

Function	Organism	Genomic localization	Advantage conferred	References
Iron uptake	Faecal *Escherichia coli*, *Klebsiella* spp., *Salmonella enterica* subgroups III and VI	Chromosome	Increased adaptability	71–73
Iron uptake, antibiotic resistance, production of lantibiotics	*Bacillus cereus*	Chromosome	Increased adaptability, competitiveness	92
Expression of adhesins	Faecal *Escherichia coli*	Chromosome	Adhesion to gastrointestinal epithelium, colonization	9
Expression of toxin complex proteins	*Photorhabdus luminescens*	Chromosome	Killing of insects, competitiveness	89,91
Expression of type III secretion system	*Bradyrhizobium japonicum*, *Mesorhizobium loti*, *Photorhabdus luminescens*, *Sodalis glossinidius*	Chromosome, plasmid	Targeted delivery of effector proteins, interference with host-cell functions	80,82,83, 86,89,91, 93,94
Expression of type IV secretion system	*Agrobacterium tumefaciens*, *Sinorhizobium meliloti*, *Bradyrhizobium japonicum*, *Mesorhizobium loti*, *Photorhabdus* spp. *Wolinella succinogenes*	Chromosome, plasmid	Targeted delivery of DNA and effector proteins into host cells, interference with host-cell functions	34,80,82, 83,86, 89–91
Nitrogen fixation, nodulation	Rhizobia	Chromosome, plasmid	Increased metabolic versatility, bacteria–host interactions	84,88
Nitrogen fixation	*Wolinella succinogenes*	Chromosome	Increased metabolic versatility	34
Sucrose uptake	*Salmonella senftenberg*	Chromosome	Increased metabolic versatility	29
Magnetic phenotype	*Magnetospirillum gryphiswaldense*	Chromosome	Magnetotaxis, increased adaptability	95
Resistance to methicillin	*Staphylococcus aureus*	Chromosome	Increased adaptability	50,51,55
Resistance to multiple antibiotics	*Shigella flexneri*, *Salmonella enterica* DT104, *Vibrio cholerae*	Chromosome	Increased adaptability	49,56,96
Resistance to mercury and kanamycin	*Providencia rettgeri*	Chromosome	Increased adaptability	58
Degradation of phenolic compounds	*Pseudomonas putida*	Chromosome	Increased adaptability	39,42
Expression of photopexin, toxins, fimbriae	*Photorhabdus luminescens*	Chromosome	Bacteria–nematode interactions, exploitation of nutrients	89–91
Expression of toxins and bacteriocins, antibiotic biosynthesis	*Photorhabdus luminescens*	Chromosome	Killing of insects, competitiveness	89–91
Myoinositol catabolism	*Photorhabdus luminescens*	Chromosome	Increased fitness, versatility	89–91
Cobalamin biosynthesis and ethanolamine catabolism	*Photorhabdus luminescens*	Chromosome	Increased fitness, versatility	89–91



Current Opinion in Genetics & Development



Pan-genome

Core-genome

Gene families shared by all genomes

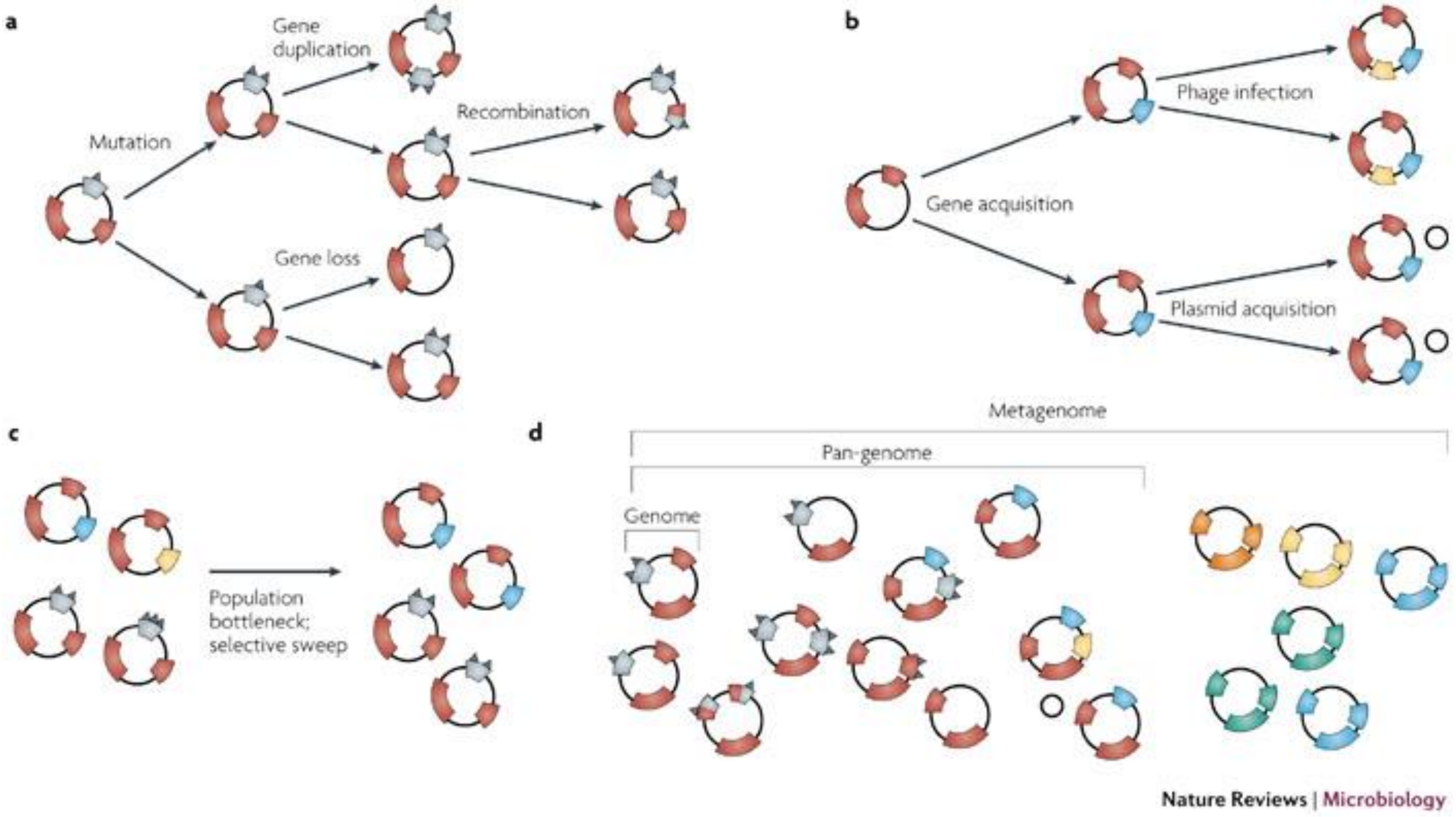
Variable-genome

Gene families shared by [2 ; n[genomes

Genome specific genes

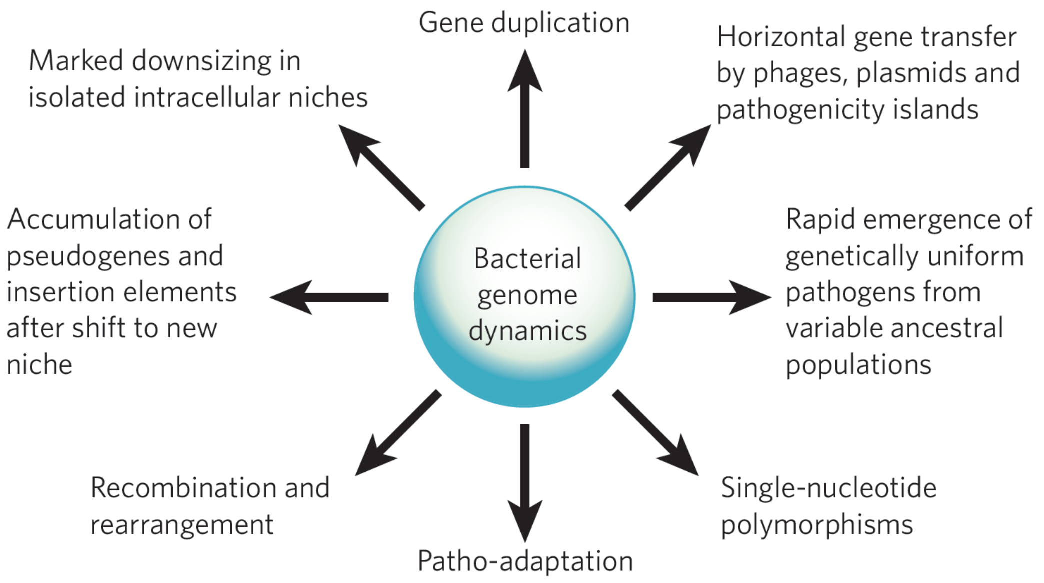
Pan-genome: The global gene repertoire of a bacterial species: core genome + dispensable genome.
Core genome: The pool of genes shared by all the strains of the same bacterial species.
Dispensable genome: The pool of genes present in some - but not all - strains of the same bacterial species.
Lateral gene transfer: Mechanism by which an individual of one species transfers genetic material (i.e. DNA) to an individual of a different species.

The set of genes pertaining to a species, or species pan-genome, depends on the number of available genome sequences. In this figure, the ***size of S. agalactiae (red dots) and B. anthracis (blue triangles)*** pan-genomes are shown as a ***function of the number of sequenced strains***. The curves represent a mathematical extrapolation of the data to a large number of strains. The **size of a species pan-genome can grow with the number of sequenced strains, or quickly saturate to a limiting value**. The S. agalactiae pan-genome is ‘open’; the B. anthracis one is ‘closed’. After sequencing a large number of strains, the number of dispensable genes in an open pan-genome is orders of magnitude larger than the size of the core genome, forcing us to reconsider the definition of a bacterial species.

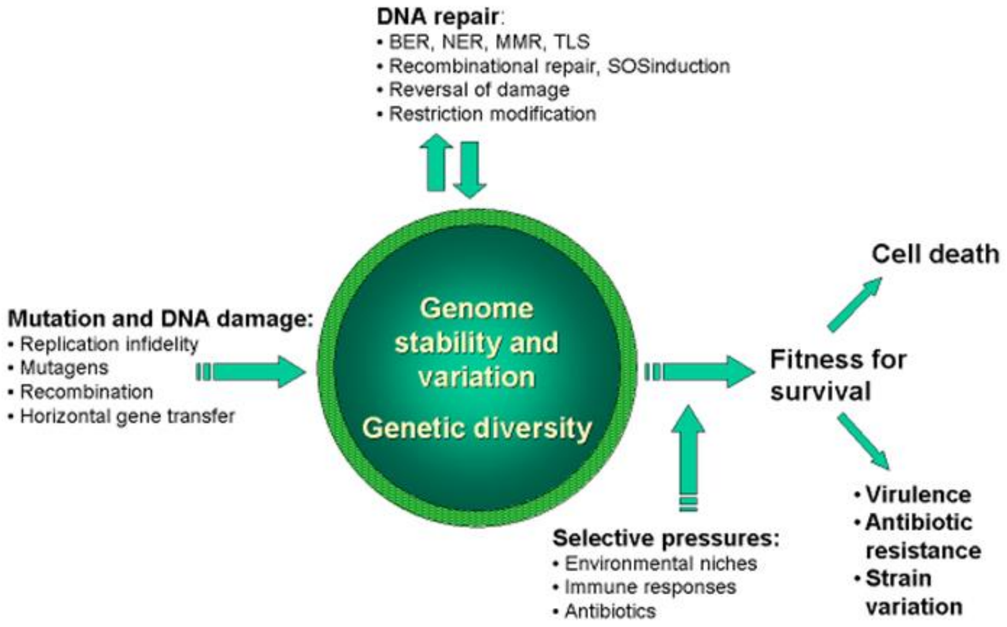


Molecular evolutionary mechanisms that shape bacterial species diversity: one genome, pan-genome and metagenome. Intra-species (a), inter-species (b) and population dynamic (c) mechanisms manipulate the genomic diversity of bacterial species. For this reason, one genome sequence is inadequate for describing the complexity of species, genera and their inter-relationships. Multiple genome sequences are needed to describe the pan-genome, which represents, with the best approximation, the genetic information of a bacterial species. Metagenomics embraces the community as the unit of study and, in a specific environmental niche, defines the metagenome of the whole microbial population (d).

Bacterial genome dynamics


Gene duplication
Horizontal gene transfer by phages, plasmids and pathogenicity islands
Marked downsizing in isolated intracellular niches
Accumulation of pseudogenes and insertion elements after shift to new niche
Bacterial genome dynamics
Rapid emergence of genetically uniform pathogens from variable ancestral populations
Recombination and rearrangement
Patho-adaptation
Single-nucleotide polymorphisms

Bacterial genome dynamics. There are **three main forces** that shape bacterial genomes: **gene gain, gene loss and gene change**. All three of these can take place in a single bacterium. Some of the changes that ***result from the interplay of these forces are shown.***


DNA repair:
• BER, NER, MMR, TLS
• Recombinational repair, SOSinduction
• Reversal of damage
• Restriction modification
Mutation and DNA damage:
• Replication infidelity
• Mutagens
• Recombination
• Horizontal gene transfer
Genome stability and variation
Genetic diversity
Cell death
Fitness for survival
• Virulence
• Antibiotic resistance
• Strain variation
Selective pressures:
• Environmental niches
• Immune responses
• Antibiotics