Genomes, genes and junk: the large-scale organization of plant chromosomes

Thomas Schmidt and J.S. Heslop-Harrison

Plants from wide taxonomic groupings have similar genes and ordering of genes along the chromosomes. However, the repetitive DNA, much of no known function, which constitutes the majority of the genome, varies extensively from species to species in absolute amount, sequence and dispersion pattern. Despite this, it is known that families of repeated DNA motifs each have a characteristic genomic location within a genus, and that there are different constraints on the evolution of repetitive DNA and genes.

There are now enough data about different types of repetitive DNA — from sequences, Southern analysis and in situ hybridization — to build a model of the organization of a typical plant genome, and apply it to gene cloning, evolutionary studies and gene transfer.

Species from wide taxonomic groupings have similar genes and arrangements of genes along the chromosomes — they show conserved syntenies. However, knowledge of syntenies — provided by high-density, marker-saturated genetic maps and genomic DNA sequence data — tells us relatively little about the large-scale physical organization of the chromosomes and the repetitive DNA elements that make up the bulk of most genomes. When a chromosome of an organism such as wheat or pine is dissected at the molecular level, stretches of nucleotide sequences that occur once or only a few times in the genome represent as little as 5% of the DNA. Most plant and animal genomes consist largely of repetitive DNA — perhaps 30 sequence motifs, typically one to 10,000 nucleotides long, present many hundreds or thousands of times in the genome — which may be located at a few defined chromosomal sites or widely dispersed. However, this repetitive DNA, with different selective pressures from those acting on genes and evolutionarily successful multigene modules, can show extensive differences in sequence motifs and abundance even between closely related species. The repetitive DNA in the genome is also important for evolutionary, genetic, taxonomic and applied studies.

Learning about the physical organization of genes and repetitive sequences, regarded by some as 'junk', and seeing where the sequences lie, is a critical element for understanding genome organization and evolution in plants. The approach enables data to be linked from Arabidopsis and the handful of smaller genomes for which sequencing is under way with other genomes that are too large and too numerous to sequence at the present time (Fig. 1).
A few repetitive sequences are known to have well-defined functions (Box 1). The telomeric sequences, added at the ends of most plant and animal chromosomes, allow a linear replication unit to be maintained, protect chromosome ends and overcome the 'end replication problem'. The 18S–5.8S–25S and 5S rRNA gene loci, clustered at a small number of sites, encode the structural RNA components of ribosomes. Mobile DNA sequences, such as transposons and retrotransposons (Box 1), make up a high proportion of most plant and animal genomes. A major class, the retroelements, encode the proteins necessary for their own reverse transcription and integration, and sometimes represent 50% of the genome. As a result of their transcription into RNA, reverse transcription into DNA and integration into the genome, they have a dispersed distribution along chromosomes. Notably, telomeres, rDNA and retroelement sequences are all ancient – they are found in all animals and plants, and might be considered as early derivatives of the 'RNA world' from which DNA-based organisms evolved.

**Fig. 1.** Species with contrasting genome sizes. Micrographs of metaphases at the same magnification show: (a) *Arabidopsis* (haploid genome containing about 140 Mbp of DNA; 2n = 10 chromosomes, where n is the basic chromosome number of the haploid genome); (b) sugar beet (750 Mbp; 2n = 18); and (c) pine (23 000 Mbp; 2n = 24). All three species are diploid; the differences in genome size arise largely from differences in repetitive DNA. Blue DAPI staining shows all the chromosomal DNA, and brighter bands show AT-rich regions. Scale bar represents 10 μm. *Pine micrograph kindly supplied by R. Doudrick*.

**Box 1. Survey of major DNA sequences of plant nuclear genomes**

<table>
<thead>
<tr>
<th>DNA repeats</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S rDNA genes with intragenic spacers</td>
<td>Low or single copy, including regulatory sequences and introns</td>
</tr>
<tr>
<td>18S-5S-25S rDNA genes with spacers</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Telomeres</th>
<th>Tandemly repeated DNA sequences</th>
<th>Degenerate copies</th>
<th>Dispersed DNA sequences</th>
<th>Others including transgenes and pseudogenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsatellites</td>
<td>Satellite DNA</td>
<td>Retroelements</td>
<td>Transposons</td>
<td></td>
</tr>
</tbody>
</table>

Satellite DNAs (Ref. 1) have varying monomer lengths, but 140–180 or 300–360 bp are frequent, corresponding to mono- or dinucleosomes. Microsatellites are runs of simple sequence repeats (with motifs 1–5 bp long); minisatellites have longer and more complex repeating units (up to 40 bp). Telomeric DNA (Ref. 26), consisting of conserved 7 bp repeats (CCCTAAA), is added to the chromosome termini by telomerase activity. Retroelements, which amplify and transpose via RNA intermediates, are divided into mobile sequences with long terminal repeats (LTRs) and non-LTR retrotransposons (long interspersed nuclear elements (LINEs) and the related short interspersed nuclear elements (SINEs)). Plant genomes could also contain solo-LTRs, miniature inverted-repeat transposable elements (MITEs) and virus-like sequences. Transposons move as DNA elements, and non-autonomous copies might be trans-activated by active autonomous elements. Connections between boxes indicate similarities in genome organization; for example, rDNA repeats are a special class of tandemly repeated DNA sequences. Arrows indicate dynamic changes between sequence classes – divergence and dispersion of tandem repeats, and clustering and homogenization of dispersed sequences.
Tandemly repeated sequences normally have characteristic chromosomal locations—sub-telomeric, intercalary or centromeric— with blocks of each motif present, in plants, on most or all chromosomes in the genome. Centromeric repeats are frequent, with arrays of 140–360 bp monomers often spanning more than 1 Mbp (Refs 6 and 7). It is notable that nucleotide stretches homologous to key parts of the yeast and human centromere boxes CDEIII and CENP-B (Ref. 8) can be identified in some plant sequences that locate at the centromere10,11 indicating that functional centromere motifs might soon be identified in higher plants. Many tandem repeat units have a complex structure, sometimes including simple sequence repeats12–14, resulting from rounds of rearrangement and amplification during evolution.

Localization of major repetitive DNA families to sugar beet chromosomes. Fluorescence micrographs taken using two different wavelengths of light allow in situ hybridization (red, yellow or green signals) to show the location of repetitive DNA motifs along 18 chromosomes (blue staining of all DNA with DAPI). A small number of repetitive sequence families represent the majority of the 750 Mbp of DNA in the genome. (a) pBV1, a 327 bp long satellite sequence, hybridizes (green-yellow) near all 18 centromeres in megabase-size blocks15, reflecting the different numbers of copies of the repeat motif. The centromeric sequences include a CA microsatellite, and the oligonucleotide shows strong hybridization at the same centromeric region16. (b) Another tandem repeat, pBV7 (a 160 bp long motif), hybridizes to intercalary sites on each chromosome arm (red); the tandemly repeated SS rRNA gene units (yellow-green) hybridize to one chromosome pair. (c) Mobile DNA sequences—retroelements—colonize most plant genomes and amplify via RNA intermediates. The Tbv retrotransposon family17 is scattered over the chromosomes (red), but is largely absent from centromeric satellite blocks. (d) Localization of cDNA shows multiple dots or bands of hybridization (red), particularly in subterminal regions, but not overlapping with the repetitive sequences shown in (a) and (c). The two larger bright red sites (bottom) represent the partially decondensed 18S-5.8S-25S rRNA genes, providing an internal control of hybridization. Single cDNA molecules (about 3000 bp long) would not give detectable signal under the conditions used, so other hybridization sites represent clusters of many cDNA probes in gene-rich regions.

Evolution of the chromosome
From comparative genome analysis, there is already a great deal of evidence for strong conservation of gene order—conserved synteny or collinearity of genes—over whole taxonomic families. Genetic maps based on DNA markers often show only a handful of chromosome rearrangements between species in taxonomic groups in both plants18–19 and animals. In maize, rice and sorghum, the gene sequence and order are highly conserved along 100 kb to 500 kb segments of the genomes in YAC clones, but the interspersed repetitive DNAs found between these genes are very different, making the physical distance between similar loci highly variable. Maize-based data show that genes behave in general as recombination hotspots20; large physical blocks of repetitive DNA contribute little to the recombination-based genetic length of the chromosome.

Some major features of plant genomes differ from mammalian and other animal genomes. For example, polyploidy and
Box 2. A model of a plant chromosome

Different classes of repetitive DNA show characteristic genomic distributions. Genes are clustered in discrete blocks between the various repetitive DNA motifs, each of which has a characteristic location and genomic organization. Ty1-copia and other families of long terminal repeat (LTR) retroelements could comprise 50% of the genome. The LTR retroelements are located throughout the genome with some clusters and depleted regions, perhaps a consequence of different evolutionary rates of sequence amplification or targeting of insertion sites. Simple sequence repeats or microsatellites are dispersed and present as clusters (shown only on the lower chromatid and varying between motifs), as well as occurring within arrays of larger tandem repeats. The 18S-5.8S-25S and 5S rRNA genes are clustered on one or more chromosome pairs in the genome (not shown). Most plant chromosomes have the 7 bp telomeric nucleotide sequence repeated at their ends. Within a plant genome, all chromosomes have a rather similar composition and organization of repeats because of homogenization of the genome, although copy numbers of individual repeat motifs might vary between chromosomes.

In sugar beet, where the average chromosome has about 80 Mbp of DNA, all the genomic elements illustrated have been found, and the species provides the reference upon which the model presented here is based. Wheat, rye, and barley, with chromosomes ten times larger, follow the same model: families of tandem repeats and the other sequence classes with characteristic locations have been found. In plants with small genomes, such as Arabidopsis (chromosome size 15–30 Mbp), major intercalary tandem repeats are not known, tandemly repeated DNA is clustered around the centromere and retrotransposons are more dispersed. Abbreviations: LINE, long interspersed nuclear element.

introduction are extremely widespread in plants, but do not play a significant role in mammalian genome evolution. At the level of chromosome organization, plants do not show chromosomal G- or R-bands, which are a consequence of differing DNA composition and function, respectively, along mammalian chromosomes. In mammalian cell-fusion and plant hybrid hybrids, total genomic DNA from one of the parental (or ancestral) species can be labelled and used for in situ hybridization to chromosome preparations from the hybrid. Chromosomes with the same origin as the probe are often strongly and uniformly labelled, while the other chromosomes are weakly labelled and often show gaps. These data show that the repetitive sequences making up the bulk of the DNA in each genome have species-specific variants enabling the discrimination. In mammals, probes derived from flow-sorted or microdissected chromosomes also label individual metaphase chromosomes within a chromosome set uniformly, and such probes are being used to show evolutionary conservation of large genomic segments by their hybridization to syntenic segments across species. The dispersed chromosome-region specific sequences responsible for the mammalian labelling properties seem to be unusual in plants: probes derived from sorted plant chromosomes do not label single chromosome types. Bivariant-flow karyotyping reveals another striking difference between plants and mammals: whereas each chromosome in a mammalian species shows a characteristic GC:AT nucleotide ratio, the chromosomes within each plant species so far examined (wheat, tomato, and field bean) all show similar nucleotide compositions, with the exception of the chromosomes with the GC-rich rDNA sequences. Taken together, these data indicate that repetitive DNA behaves differently in mammals and plants, with plants showing greater homogenization between all chromosomes in a species: both dispersed repeats, such as retroelements, and tandemly repeated motifs show interchromosomal homogenization in plants.

Perspectives

We expect that large-scale genome organization will follow similar principles to this model in plant species with smaller, more streamlined genomes, and in larger genomes, whether diploid or polyploid. The clustering of genes is good news for map-based gene isolation strategies, and the mosaic of repetitive sequence arrays provides physical reference points to map and order genes, localize transgenes or characterize sites of recombination.

Acknowledgements

T. Heslop-Harrison was supported by the Deutsche Forschungsgemeinschaft (Schn 1048/1-1 and Schn 1048/2-2). J.S.H-H. thanks the Biotechnology and Biological Sciences Research Council (PAGA initiative) and the EC Framework IV Biodiversity project (ERB- BIOACT19960508) for support, and the British Council/german Academic Exchange Service (DAAD) programme for facilitating joint work. We thank our colleagues for many valuable discussions.

References

5. Heslop-Harrison, J. S. et al. (1997) The chromosomal distribution of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution, Genetics 100, 197–204

Drugs from nature

Medicinal Natural Products: A Biosynthetic Approach by P.M. Drouet
John Wiley & Sons, 1997. £29.95 pbk (ix + 466 pages) ISBN 0 471 57478 1

Natural product drugs have been used for centuries for therapeutic purposes. They were used initially as the whole herbs or extracts, and later as the purified natural products (or as closely related chemical derivatives). Plants have the longest history of use, although more recently animals (mostly marine) and particularly microorganisms have also been used extensively as sources of drugs. Research into natural product drug discovery has been revitalized by recent findings of important natural products such as taxol, podophyllotoxin and artemisinin. The recent emergence of herbal remedies as alternatives to synthetic drugs has also generated much interest in the field.

This book is a well-written description of current information on medicinal plants. The emphasis of the book is on biosynthesis – the way that secondary metabolites of medicinal value are synthesized from simple precursors. These biosynthetic relationships are covered comprehensively – all of the major biosynthetic schemes for natural products are illustrated, and the examples have been well chosen. Fundamental aspects of secondary metabolism are covered in a general chapter, and this is followed by chapters that each address specific pathways/classes (e.g. acetate, shikimate, mevalonate, alkaloids, peptides and carbohydrates). The majority of the text deals with the biosynthetic relationships involved in explaining the biosynthetic pathways. The diverse structures are clearly drawn, with the stereochemistry and the numbering systems indicated on key structures. The author also chose to make ‘inserts’ or monographs at various places that describe the roles of the natural products as drugs.

These inserts include brief, practical information on, for example, sources, uses and biological properties.

Medicinal Natural Products is primarily targeted as a textbook for undergraduate pharmacy students. The general organization is student-friendly, and the first chapter on how to use the book is helpful. The table of contents and index also make finding information relatively easy. Nevertheless, given that many pharmacy courses in the USA have reduced the pharmacognosy element in their curricula in the past 10–15 years, it is a pity that more space is not given to information relevant to the practice of pharmacy – this is confined to the ‘monograph’ element of the book. The division of the subject material by biosynthetic classes, rather than biological class or therapeutic activity, means that the book cannot be recommended as a textbook for undergraduate pharmacy students in the USA. I believe that it is better suited for biosynthetic courses in graduate studies and/or specialized undergraduate projects covering natural products – it would be especially useful for students in pharmacy graduate programs who...