Review TRENDS in Genetics Vol.22 No.3 March 2006

Diagnosing duplications – can it be done?

Dannie Durand[1] and Rose Hoberman[2]

[1]Departments of Biological Sciences and Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

New genes arise through duplication and modification of DNA sequences on a range of scales: single gene duplication, duplication of large chromosomal fragments and whole-genome duplication. Each duplication mechanism has specific characteristics that influence the fate of the resulting duplicates, such as the size of the duplicated fragment, the potential for dosage imbalance, the preservation or disruption of regulatory control and genomic context. The ability to diagnose or identify the mechanism that produced a pair of paralogs has the potential to increase our ability to reconstruct evolutionary history, to understand the processes that govern genome evolution and to make functional predictions based on paralogy. The recent availability of large amounts of whole-genome sequence, often from several closely related species, has stimulated a wealth of new computational methods to diagnose gene duplications.

Introduction

The modern genome is a palimpsest (see Glossary) of small and large-scale duplication, genome rearrangement, gene loss and sequence divergence. The challenge in diagnosing duplications is to interpret superimposed evidence of small-scale (e.g. tandem duplication and retrotransposition) and large-scale duplication (block duplication, aneuploidy and whole-genome duplication) in the face of rearrangement, loss and mutation. Accurate diagnosis is difficult, as can be seen from the conflicting results of different analyses of the same data sets (Box 1 and Refs [1–4]). These difficulties have provoked an emerging awareness of the need for formal methods to diagnose duplications. Early studies were based on simple, *ad hoc* heuristics, whereas studies that have appeared more recently use modeling and statistical validation. In this article, we review the computational methods for analyzing spatial and temporal data to diagnose gene duplication processes. Novel genes arising through domain shuffling and/or duplication of gene fragments [5] are beyond the scope of this article, as are duplicated non-coding sequences [6].

The availability of new data and experimental methods for investigating duplications in the laboratory has led to myriad studies investigating the history, type and fate of duplicated genes in model organisms [7–10]. Recently, there has been particular interest in how the fate of duplicated genes depends on the duplication mechanism [11–13]. Duplicated genes are classified into a few groups, characterized by the spatial signature associated with a specific duplication mechanism. Methods to identify these signatures can be either map- or sequence-based. Map-based approaches treat the genome as an ordered list of genes (the map), ignoring intergenic sequence. A set of potential paralogs is identified by selecting significantly similar gene pairs from an all-against-all comparison of the gene sequences (or their products). Additional filtering can be imposed at this stage to remove spurious matches (e.g. matches as a result of shared domains). Once homology relationships between the sequences have been established, local similarities in gene order and content are sought by comparing the map with itself. Distinct regions with many shared paralogous genes are candidate duplicated regions. The challenge is to determine whether these putative duplicated regions are the result of fragmented whole-genome duplication (WGD), other large-scale duplication processes or a cluster of single gene duplications (SGDs) located near each other by chance.

Glossary

Allopolyploidy: the creation of a genome with doubled chromosome number through fusion of cells from two closely related species with the same ploidy.
Aneuploidy: the presence of extra copies, or no copies, of some chromosomes.
Autopolyploidy: doubling every set of homologous chromosomes in the genome.
Greedy algorithm: an algorithm that at each stage makes the locally optimal choice. Depending on the nature of the problem, this strategy might not produce a globally optimal result.
Large-scale duplication: whole genome, chromosomal or block duplication.
Non-allelic homologous recombination (NAHR): crossing over mediated by DNA mispairing, resulting in duplication and/or deletion of DNA fragments.
Palimpsest: a parchment that has been written on and partially erased repeatedly, so that ancient and recent messages are intermingled.
Paralog: homologous genes related by duplication.
Paralogon: putative duplicated blocks, characterized by pairs of non-overlapping chromosomal regions, enriched for paralogous gene pairs.
Polyploidy: whole genome duplication.
Segmental duplication: duplicated regions that are 1–20 kb in length with at least 90% sequence identity, also known as low-copy repeats. The term 'segmental duplication' has also been used to describe any duplication of a large chromosomal fragment in a single event and any duplication that can not be shown to be the result of a tandem or large-scale duplication.
Single gene duplication: a tandem duplication or retrotransposition.

Corresponding author: Durand, D. (durand@cmu.edu).
Available online 26 January 2006

 0168-9525/$ - see front matter doi:10.1016/j.tig.2006.01.002

Box 1. Case study: diagnosing large-scale duplication in rice

The rice self-comparison map, which is 'noisy' compared with the self-comparison maps of other putative polyploids, presents a particular challenge for duplication diagnosis. It also shows that the presentation of similar duplication processes can differ significantly in different lineages. Thus, methods of duplication analysis that were developed for one genome might yield misleading results if applied to another genome without modifications.

Rice is characterized by many randomly distributed paralogs that obscure the regular patterns associated with both tandem and large-scale duplication. To find paralogons in rice, the 'noise' that obscures regular patterns of large-scale duplication in the dot-plot must be addressed. Two groups filtered paralogs in a preprocessing step to enable these patterns to emerge, either by restricting paralog data to gene families that contain two members [62] or to paralogs with K_s values of 0.75–0.95 [46]. In both cases, filtering reduced the randomly distributed matches and enabled paralogon detection in the remaining data using a relatively conservative paralogon definition or simply by inspection. The results support a polyploid origin for rice, with at least one subsequent block or chromosomal duplication [45,46,61,62]. The paralogons identified by these studies were distributed across all twelve chromosomes and covered 45–62% of the genome.

By contrast, Vandepoele and colleagues [59,60], using a conservative paralogon definition requiring conserved order, identified paralogons that covered $\leq$20% of the rice genome, primarily on chromosomes 2, 11 and 12. Based on the low coverage and non-uniform distribution observed, they argued that the spatial evidence supports aneuploidy rather than polyploidy. The use of a conservative definition, without prefiltering, might explain the lower coverage numbers reported. Although in *Arabidopsis* this conservative definition detected most duplicated blocks, the noisier nature of the rice data set might require a more flexible definition [45].

The atypical pattern of duplication in rice also obscures patterns of large-scale duplication in temporal data. Histograms of rice paralogs exhibit a large peak at low K_s values, and a low, broad bulge at K_s= 0.8 [25,59]. This secondary peak is partially obscured by the abundance of SGDs, leading some authors to argue that the peak reflects aneuploidy rather than polyploidy [59]. This view was influenced by results from *Arabidopsis*, because the peak in rice is much smaller than the peak associated with the postulated WGD in that species. However, recent spatial analyses contradict the hypothesis of a single aneuploidy event in rice. Thus, diagnosing large-scale duplication in rice is difficult because empirical distributions resulting from WGDs can vary substantially from one lineage to the next.

Recent tandem duplications are easy to detect with map-based approaches because they occur in arrays of two or more genes with significant sequence similarity found in close proximity on the same chromosome. These tandem duplications are thought to be duplicated by non-allelic homologous recombination (NAHR) mediated by alignment of mispaired repetitive sequences, although other mechanisms have been proposed [14]. Block duplication – duplication of large chromosomal fragments – can occur through DNA transposition or translocation followed by meiosis.* These were first observed in the context of cytogenetic studies of karyotypic abnormalities [15]. In comparative mapping, block duplications manifest themselves as regions enriched for paralogous pairs in genome self-comparisons. Duplication of individual chromosomes (aneuploidy) can also be observed in this manner.

Whole-genome duplication (WGD) can arise through either autopolyploidy, doubling every set of homologous chromosomes in the genome, or allopolyploidy, creation of a genome with doubled chromosome number through interspecific hybridization. Recent polyploidy can be inferred through direct observation of multivalent synapsis or by comparing chromosome numbers in closely related species. Similarity in genomic organization is a source of evidence for inferring ancient polyploidy in species that have reverted to diploid segregation. However, regions of similarity will be degraded by subsequent gene duplication and loss, and fragmented by large-scale rearrangements, making it difficult to distinguish WGDs from other large-scale events.

Although a large proportion of paralogs in the genome can be diagnosed using map-based approaches, there will be others that have an unclear origin. These include older tandem and large-scale duplications that have been dispersed by subsequent rearrangement, so that their characteristic spatial patterns are no longer apparent. Temporal analyses, based on estimates of duplication times, can help to identify the origin of some of these paralogs; for example, a preponderance of duplicates of the same age suggests a WGD occurred.

Neither method will identify retrogenes, genes that are duplicated by reverse transcription of an mRNA followed by reinsertion of the resulting cDNA into the genomic sequence. This process is thought to be mediated by reverse transcriptases that are carried by transposable elements (TEs). Because the retroposed mRNAs are usually devoid of regulatory sequences, retrogenes do not represent a large proportion of functional paralogs in most genomes, although a few functional retrogenes have been reported [16,17]. Although retrogenes can be recognized by their lack of introns and, in some cases, the presence of poly-A tails and flanking repeat sequences associated with the integration sites of TEs [18], this analysis is rarely included in map-based studies.

Sequence-based methods have proved most useful for identifying retrogenes [18] and low-copy repeats (LCRs), recent highly conserved duplications [19] with an unknown mechanism of formation (Box 2). In studies of recent duplications, sequence analysis has also been used to diagnose the genomic mechanisms that drive various duplication processes [20,21]. Evidence of transposition can be found in the characteristic sequence patterns associated with known families of TEs and their integration sites. NAHR, which mediates tandem duplication and, in some cases, translocation, is characterized by repetitive sequences of the same type in both regions that flank a duplication. Preservation of order and orientation with respect to the centromere and/or telomere is also evidence of translocation. Although certain studies [1,22] have attempted to use genome-scale sequence comparison to diagnose whole-genome duplications, homologous regions in ancient polyploid genomes are typically too diverged to permit sequence alignment outside of coding regions.

The challenges in diagnosing duplications are compounded by substantial, lineage-specific differences in the frequency and type of duplications that occur. Recent

* Block duplications are sometimes referred to as 'segmental duplications' but this term has also been used to describe LCRs (Box 2). Currently, the evidence is not sufficient to determine whether these are the same or different phenomena.

Box 2. Low-copy repeats

LCRs are highly conserved, duplicated regions, defined operationally in terms of their length (1–20 kb) and degree of sequence conservation (90–99.5%) [19]. The upper limit on sequence identity is imposed to avoid confusing true duplications with assembly errors, whereas the minimum length requirement screens out repeat sequences that are associated with transposable elements. Estimates of the contribution of LCRs to genome sequence are sensitive to assembly, coverage, allelic variation and annotation of repetitive sequences.

Because reports conflict on the gene content of LCRs [20,21,67–70], depending on methodology, species and the quality of the data, it is not clear to what extent LCRs have a role in the duplication of entire genes. However, there is substantial evidence that LCRs contain fragments of coding sequence and are likely to be the sites of new gene formation by domain shuffling.

The origins of LCRs are not well understood. Proposed mechanisms include translocation followed by transmission of unbalanced chromosomal complements in human subtelomeric regions [20], *Alu*-mediated transposition in pericentromeric regions [21,71,72], copy number expansion, owing to NAHR mediated by DNA repeats, and chromosomal instability owing to variations in supercoiling [73]. Duplication activity is characterized by short intense periods of activity, followed by quiescence. Moreover, patterns of duplication differ from lineage to lineage.

The contribution of LCRs to total genomic sequence also varies substantially between lineages, with more LCRs in human and chimpanzee than in other species. The disparity between lineages is probably due to specific characteristics of genome structure and dynamics [72]. These include sub-terminal caps, which seems to be an innovation that is specific to great ape genomes, variations in TE composition and transposition rates, and the emergence of α-satellite repeats.

Given the obstacles to studying LCRs and the uncertain nature of the current evidence, it is difficult to determine whether LCRs are the most recent and highly conserved examples of a continuous duplication process or a phenomenon of particular importance in primates engendered by changes in genome structure and dynamics. At this point, it is unclear whether LCRs should be considered among the possible alternate hypotheses for the origin of putative duplicated regions when analyzing ancient polyploids.

studies have focused on elucidating the main principles of genome evolution with less attention to how those principles are modulated by lineage-specific features. However, the evidence suggests that different duplication processes dominate in different lineages. The genomic distribution of TEs and rates of proliferation and loss vary greatly with lineage [23]. This has a direct effect on duplication processes mediated by transposition and an indirect effect on processes that are mediated by the presence of repetitive elements, such as NAHR. This is consistent with reports that rates of tandem duplication differ significantly from one lineage to another [24,25].

The forces of selection that influence the persistence of duplications once they arise are also lineage specific. Aneuploidy is considered to be a rare event in the evolution of animals because of the associated deleterious dosage effects, but is more often tolerated in plants [26] and yeast [14,27]. Although recent studies have presented evidence of polyploidy in early vertebrates [28], fish [29–31], yeast [32,33] and numerous plant species [34], polyploidy does not seem to be a universal evolutionary process: large-scale duplication is not observed in fly, worm or bacteria. Forces believed to modulate the survival of newly arisen polyploids included the propensity of genome duplication to disrupt the genetics of sex determination, physiological constraints imposed by the impact of polyploidy on cell size and the difficulty of finding a genetically compatible mate. These observations led to the prediction that polyploidization will rarely be observed in animals but more frequently in plants. However, with the discovery of many additional polyploid species [35], the emergence of molecular methods in the 1980s and 1990s and the increasing availability of genomic sequence, much of the accepted pre-genomic lore concerning polyploids has been questioned [36–38]. This has led to a resurgence of studies of the mechanisms of polyploidization, the physiological conditions that enable polyploids to arise, positive and negative forces acting on neopolyploids and the extent to which polyploidization represents special evolutionary opportunities.

Temporal analysis

A large-scale duplication, whether a whole-genome, chromosome, or block duplication, will result in a preponderance of paralogs that originated simultaneously; therefore, the age distribution of duplicated genes can be used to diagnose large-scale duplications. A common approach is to plot a histogram of the number of duplicated genes against estimated duplication times. A large-scale duplication will be visible as a peak in this distribution, superimposed on a background of ongoing SGD (Figure 1). To tease the different duplication processes apart, it is necessary to distinguish the characteristic distributions of each process, taking distortions caused by inaccuracy in time-estimation methods into account.

Duplication times of individual gene pairs are estimated by a variety of methods (reviewed in Ref. [7]). Most

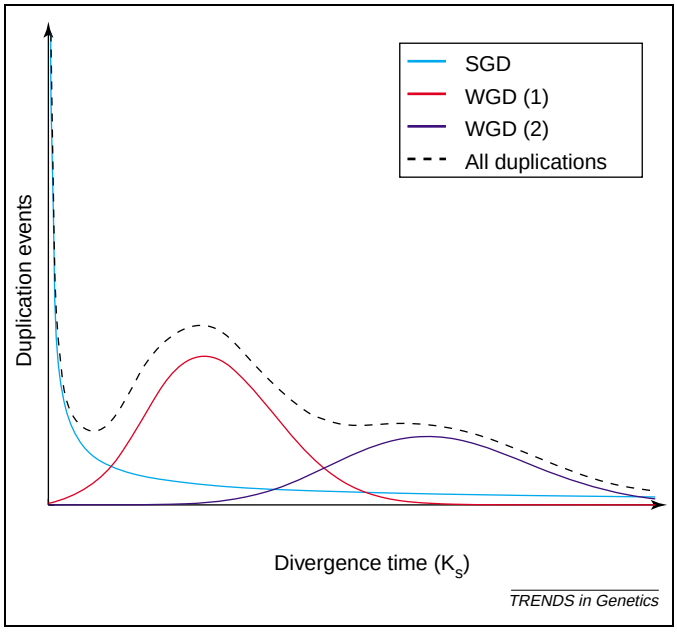

Figure 1. A theoretical age distribution of pairs of duplicated genes in which age is expressed as the proportion of synonymous substitutions per synonymous site (K_s). A complex distribution can result from the superposition of several different duplication processes, in this example two whole-genome duplications (WGD1 and WGD 2) and ongoing small gene duplication (i.e. SGD).

commonly, the fraction of synonymous substitutions per synonymous site (K_s) is used as a relative measure of the time that has elapsed since duplication. Time estimates will vary owing to the inaccuracy in corrections for multiple substitutions and variation in substitution rates [39]. Moreover, the molecular clock is 'sloppy' – even when the average number of substitutions is approximately linear with time, the variance will be significant [39]. Thus, a large-scale duplication will be visible as a broad peak in the distribution of duplication times, rather than a narrow spike.

It is possible to detect recent WGDs in the histogram by inspection (e.g. Refs [3,40]). For example, Blanc and Wolfe constructed histograms for 14 model plant species, and argued that nine of the species have 'age distributions of paralogous genes that are incompatible with a null model of gradual gene duplication and loss but similar to what is expected from large-scale duplications such as polyploidy or aneuploidy' [25].

For older events, identifying a WGD against a background of ongoing SGD is more difficult. Recent efforts have moved towards developing formal mathematical models of gene duplication. Typically, a parametric distribution is selected to model duplication and loss rates of a particular duplication process. The parameter values are determined by fitting the model to the data. A statistical test is often conducted to determine whether the data could have been generated by the proposed model. Different hypotheses can be compared by testing whether one model fits the data significantly better than another, taking into account the number of parameters in each model.

Lynch and Conery [41,42] were among the first to fit parametric distributions to K_s histograms for duplicated genes, although not in the context of diagnosing duplication. Based on the shape of the histograms obtained, they postulated that the age distribution of duplicated genes could be explained by a birth–death process (Box 3). Under this model, they estimated duplication and loss rates in each species but did not formally test how well the model fit the data. More recently, statistical tests such as the parametric bootstrap have been used to determine whether the histogram of duplication ages in humans can be explained by a birth–death process [43,44].

Maere *et al.* [13] took this approach several steps further, with a formal model of both single gene duplication (SGD) and WGD in *Arabidopsis*. The model incorporates a continuous, constant-rate process of SGD and three discrete WGD events, each associated with a different time-dependent rate of loss (Box 3). In addition, the dispersion of K_s values that arises owing to the sloppy nature of the molecular clock is modeled explicitly using a Poisson distribution.

The resulting multi-component model is used to test various alternate hypotheses concerning the relative importance of SGDs and WGDs in the evolution of the *Arabidopsis* genome. By visually comparing the fit of different model variants with the actual data, they draw several specific conclusions, including: (i) three WGDs are more consistent with the data than two; (ii) gene loss fits a power law rather than an exponential distribution; and (iii) for most functional categories, the loss rate after SGD is inversely correlated to the loss rate after WGD. Visual analyses, however, can be misleading. For example, when comparing models with different numbers of parameters, a better fit could simply be due to the additional parameters. In such cases, formal statistical tests of goodness-of-fit are crucial.

The power of highly detailed, parametric modeling derives from the ability to combine multiple processes in a single model and use rigorous statistical approaches to test hypotheses concerning evolutionary events. However, the complexity that gives this approach its strength is also a potential pitfall. Conclusions drawn from models are only as trustworthy as the assumptions on which the models are based. For example, an incorrect number of WGDs can be inferred in examples where a WGD is not characterized by a single unimodal peak. If the WGD was an allopolyploidy and diploidization did not happen immediately, then a bimodal K_s distribution could result (one peak corresponding to the time that has elapsed since the WGD event occurred and another to the time that has elapsed since rediploidization) [10]. Bimodality can also result from variations in GC content [45]. Furthermore, with a complex model of multiple processes, hypotheses about one process rest on the assumption that the models of the other

Box 3. The age distribution of single gene duplications

A common model of single gene duplication is a birth–death process, characterized by a constant birth rate, μ, and a constant death rate, λ. In this model, a histogram of the K_s values of duplicated genes will display an L-shaped distribution with a peak near zero and an exponentially decreasing tail (shown as a black line in Figure Ia). The shape of this curve can be understood by observing that although most recently duplicated genes will be apparent, many of the ancient duplications will have been lost or obscured by mutation [42].

A gene that has been retained for a significant period of time is probably under selective pressure and thus less likely to be lost, a phenomenon not captured by the birth–death model [13,44,74]. This pattern of loss can be modeled with a time-dependent decay rate λ/t, which decreases as the retention time increases. Unlike the exponential decay resulting from a constant-rate birth–death process, this model leads to a power law decay with a thicker tail [13] (shown as a blue line Figure Ia).

Both of these models treat SGD as a single, continuous process with an L-shaped distribution that peaks at $K_s = 0$. These models fit empirical histograms well in some examples and poorly in others. SGD in barley, for example, display the expected L-shaped distribution (shown as a red line in Figure Ia). In rice and *Arabidopsis,* however, histograms of tandem duplications (identified by their spatial arrangement) exhibit a poorer fit. The distribution of tandems in rice peaks at $K_s = 0$, but exhibits a slight, unexpected, peak near $K_s = 0.4$ (Figure Ib). In *Arabidopsis*, the tandem histogram differs from the theoretical distributions owing to a trough near $K_s = 0.1$, followed by a striking secondary peak at $K_s = \approx 0.4$ (Figure Ic).

The second peak can not be explained by a single large-scale event because there are no known mechanisms that can cause multiple, simultaneous tandem duplications at disparate loci. More-realistic alternative hypotheses are also more complex: the peak could be caused by a sudden increase in the rate of tandem duplication, or the trough could reveal a recent increase in the rate of DNA deletion in *Arabidopsis*, with young tandems that have not yet diverged significantly in function being preferentially deleted [25]. Because there is evidence suggesting that the *Arabidopsis* genome shrunk significantly during the past ~50 million years [25], the second explanation seems more probable.

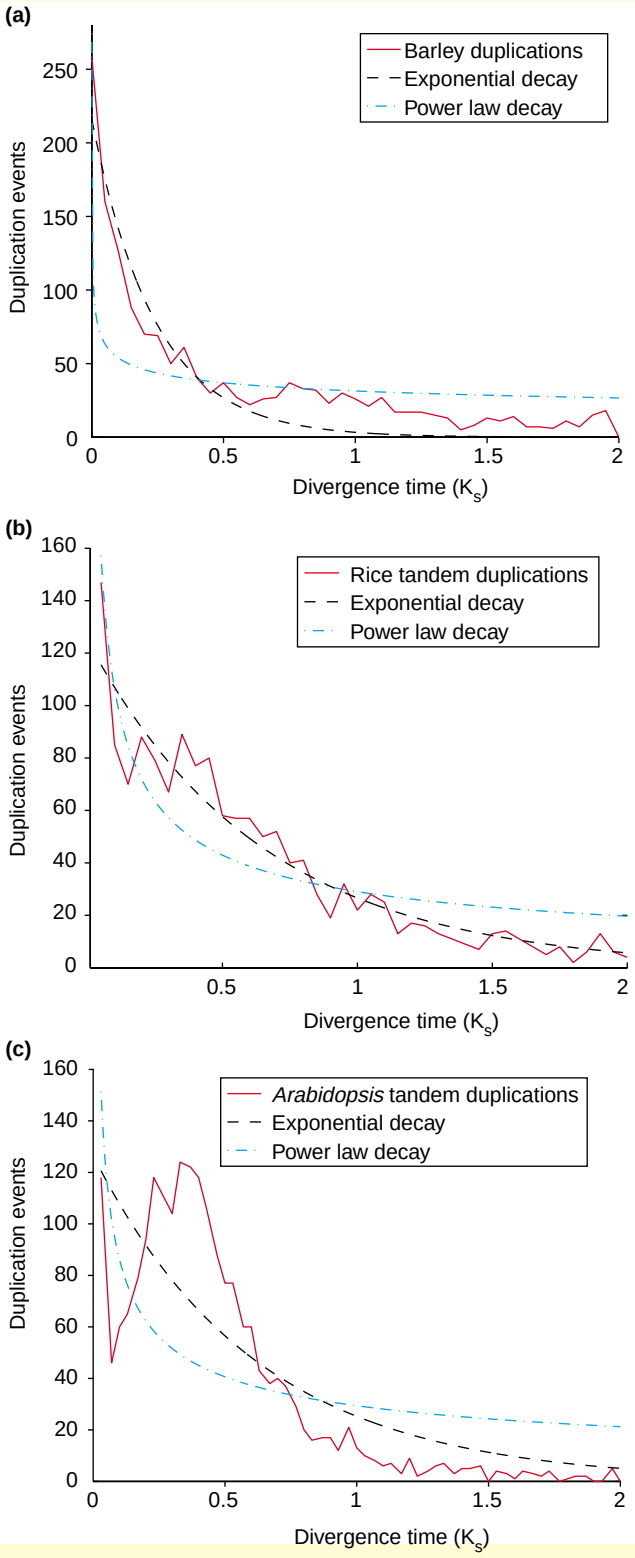

Figure I. Comparing theoretical models (exponential and power law decay) of single gene duplication with empirical K_s distributions of **(a)** all duplicated genes in barley; **(b)** tandem duplications in rice; and **(c)** tandem duplications in *Arabidopsis*. Theoretical distributions were fit to the empirical data by minimizing the weighted sum of the squared distances. This figure is reproduced, with permission, from Ref. [25] ©The American Society of Plant Biologists.

processes are correct, leading to circular reasoning. For example, the tandem duplications shown in Figure Ic in Box 3 do not fit the theoretical models. The secondary peak in *Arabidopsis* tandems could be mistaken for a WGD in the absence of spatial data to confirm tandem duplication as the mechanism of origin [25].

Spatial analysis of large-scale duplications

Large-scale duplications will be evident as regions in the genome that have similar gene content and, perhaps, similar order. Fragmentation, rearrangement, gene loss and subsequent SGDs will degrade these regions, obscuring the evidence of large-scale duplication.

The key step in spatial diagnosis of large-scale duplications is the identification of putative duplicated blocks. Map-based approaches start by identifying paralogs potentially duplicated in a single event. If a particular large-scale event is the focus of inquiry, temporal or phylogenetic evidence can be used to restrict the set of paralogs under consideration to those duplicated in the relevant time frame (e.g. Refs [3,46]).

Recognizable tandem arrays, defined to be a series of paralogs that are in proximity to each other on the same chromosome, are 'filtered out' by collapsing each array into a single representative gene. To account for subsequent rearrangements and insertions, many studies permit a limited number of unmatched genes (typically, one to 30) between any pair of paralogs on the same chromosome [1,3,28]. Other studies require that a tandem array containing N paralogs spans no more than $2N$ genes in total [47], or restrict a tandem array to paralogs that are on the same bacterial artificial chromosome (BAC) [48].

The filtered set of paralogs is used to identify putative duplicated blocks, or paralogons. Various approaches have been proposed for identifying paralogons based on map self-comparison [49], and, in some cases, this is combined with a comparison with a pre-duplication species [29,50–53]. Proposed paralogons are then refined by statistical analysis to rule out the null hypothesis: that the cluster of paralogs resulted from several independent SGDs that were inserted in the same region by chance. This is most commonly achieved by randomization, although formal statistical methods are beginning to emerge [54–58].

Finally, the resulting set of paralogons is analyzed to determine the number, type and timing of large-scale duplications, and to determine whether individual gene duplications are the result of a block, chromosomal or whole-genome duplication. Except in rare cases, such analyses are not based on formal hypothesis testing. When the genome sequence of a closely related pre-duplication species is available, it is relatively straightforward to distinguish WGD from block and chromosomal duplications. If all paralogons correspond to regions on a single chromosome in the related species, the evidence suggests a fragmented aneuploidy. If each region in the genome of the pre-duplication species maps to two paralogons, with no overlaps, a WGD occurred. If more than two paralogons map to one region, then repeated block duplications are indicated; but if there are several paralogons in the region it is possible that a second WGD occurred.

When no appropriate pre-genome duplication is available, temporal and spatial features of paralogons are analyzed. Paralogons resulting from a WGD are expected to have similar estimated duplication times, to not

overlap, to cover a significant portion of the genome and be uniformly distributed across the chromosomes [28]. Paralogons that cover only a small portion of the genome or have skewed spatial distributions are more likely to have arisen by block duplication. Even when the coverage and distribution of paralogons suggest a WGD, additional large-scale duplications might be discovered by analyzing duplication times. If the estimated age of one paralogon is radically different than the ages of the bulk of paralogons, it is likely to have arisen by a block or chromosomal duplication. In addition, the presence of overlapping paralogons suggest multiple large-scale events, although when there are only a few overlapping regions, it might be difficult to determine whether the older event was an ancient WGD or multiple block duplications.

Paralogon detection

To identify ancient duplicated regions, it is necessary to define the spatial patterns suggestive of common ancestry and then design a search algorithm to find such patterns. However, published paralogon definitions differ substantially. Consequently, when different groups analyze the same data set they often obtain different results. For example, in rice, paralogons identified by different methods comprise different proportions of the genome, ranging from 15% to 66% [45,46,59–62].

Although some of the disparity can be attributed to the considerable variance in homology identification, much of the variation is due to different strategies for paralogon detection and validation. As a community, we share an intuitive notion that ancient duplicated regions will be enriched for homologous gene pairs, but that neither gene content nor order in these regions will be strictly preserved. However, there are no accepted, formal specifications for such regions, nor is there an explicit discussion of the properties that are desirable in such a specification. As a first step, we have proposed properties whereby paralogon definitions can be compared and evaluated [63]:

(i) Order: a paralogon is completely ordered if the order of the paralogs in one region is either identical to or the exact inverse of the order in the other region.
(ii) Nestedness: a paralogon of size k is nested if it contains within it a paralogon of every size $<k$ (i.e. of size 1, 2, …, and $k-1$).
(iii) Isolation: a pair of paralogons is isolated if the maximum distance between adjacent paralogs in either paralogon is less than the minimum distance between the paralogons.
(iv) Size: the number of paralogs contained in a paralogon.
(v) Density: paralogon size divided by length, where length is defined to be the total number of genes, even if they are not paralogous, contained in a paralogon.

We illustrate these properties with the widely used max-gap paralogon definition (reviewed in Ref. [57]). A max-gap paralogon is a maximal set of paralogs in which the number of interlopers between adjacent paralogs is never greater than some specified number, d. For example, when $d=1$ the chromosomal regions in Figure 2 contain three paralogons. When $d=0$, only one paralogon is found, containing the four leftmost paralogous pairs in P2.

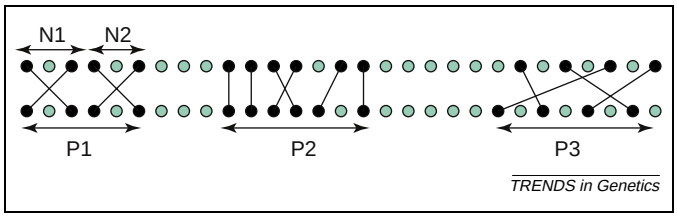

Figure 2. A genome self-comparison. Paralogs are shown in black and are connected by lines. The genes shown in green are unmatched in the visible region but can be paralogous to genes elsewhere in the genome. If one gene, at most, is permitted between paralogs ($d=1$), three paralogons are found: P1 (size four; containing four paralogs, density 4/6), P2 (size six, density 6/7) and P3 (size four, density 4/7). P1 and P3 are not nested paralogons, so if the search is restricted to nested paralogons only N1, N2 and P2 are found.

The properties were selected to evaluate the ability of paralogon definitions to capture properties of biological interest. Regions that were duplicated recently will be characterized by conserved order. Relaxed order constraints enable the recognition of duplicated regions that have been partially scrambled by subsequent rearrangement. Differences in order constraints imposed in each study might explain much of the inconsistency in rice coverage numbers (Box 1).

Even if a definition seems not to place any constraints on order, the search procedure to find paralogons can constrain order implicitly, a phenomenon captured by the nestedness property. Nested paralogons are inherently more ordered than unnested paralogons (e.g. compare P2 and P3 in Figure 2). 'Greedy' search methods are only guaranteed to find nested paralogons, because they construct larger paralogons from smaller ones. Such agglomerative procedures are widely used (cited in Ref. [57]), but will not find highly disordered clusters. For example, P3 (Figure 2) would be missed by a greedy strategy because it does not contain any paralogon that contains two paralogs when $d=1$ [57,64]. Therefore, studies that use 'greedy' methods are implicitly imposing order constraints on the paralogons identified, and thus might inadvertently fail to discover highly disordered duplicated blocks. Although they are not currently in widespread use, divisive top–down methods have been developed that will detect both nested and nonnested paralogons [64].

The isolation property expresses the expectation that duplicated regions will be islands of paralogy separated by seas of interlopers. Many definitions, however, result in distinct paralogons that are in close proximity. For example, requiring paralogons to be nested can lead to a somewhat unsatisfactory scenario in which the distance between two paralogons is less than the distance between paralogs within a paralogon. In Figure 2 P1 is not nested; a strictly greedy search would not find P1 but would identify only its two sub-paralogons, N1 and N2, which are nested. However, it seems more natural to consider the

four paralogs as a single duplicated block because N1 and N2 are adjacent.

Candidate paralogons are often evaluated based on their size or density. Some studies define paralogons in terms of a minimum size requirement, whereas others constrain the density, resulting in different sets of paralogons. For example, McLysaght *et al.* [3] searched for dense paralogons in the human genome by constraining the maximum number of interlopers permitted between adjacent paralogs in each paralogon to $d=30$. Based on randomization tests, they determined that any paralogon containing six or more paralogs was statistically significant (i.e. was likely to have been formed by a single, large-scale duplication). By contrast, in a separate analysis of human duplications, Panopoulou *et al.* [4] constrained the size of the paralogon to contain at least two paralogs, then used randomization to find the maximum number of interlopers that would result in statistically significant paralogons ($d=10$).

The results of the two approaches differed significantly in terms of number of genes found in paralogons and the distribution of paralogon sizes. Most of the paralogons identified by Panopoulou *et al.* [4] contained two or three paralogs. The length of the largest paralogon was six and contained five paralogs. By contrast, the significant paralogons identified by McLysaght *et al.* [3] contained at least six paralogs; the length of the largest was 63 and it contained 29 paralogs. These differences arose because each group evaluated paralogons based on only one property (size or density), rather than considering the interaction between the two properties.

Paralogon size is often used as a test statistic in statistical tests, reflecting the assumption that although small groupings of paralogs can occur by chance, any paralogon with a substantial number of paralogs must have arisen through a large-scale duplication [7]. However, formal statistical analysis reveals that these intuitive notions are not always valid. For definitions that constrain only the number of interlopers, for example, the probability of observing a paralogon by chance can actually increase with the size of the paralogon [57] (Figure 3). This observation has implications for the choice of test statistic. In a standard hypothesis test, the *P*-value is defined as the probability under the null hypothesis of obtaining a value of the test statistic that is at least as extreme as the observed value. However, when a larger paralogon is actually more likely to occur by chance, a larger value of the test statistic is not more 'extreme' from a statistical viewpoint. This is not merely an abstract statistical issue but suggests a failure to capture the full interaction between paralogon properties and paralogon significance.

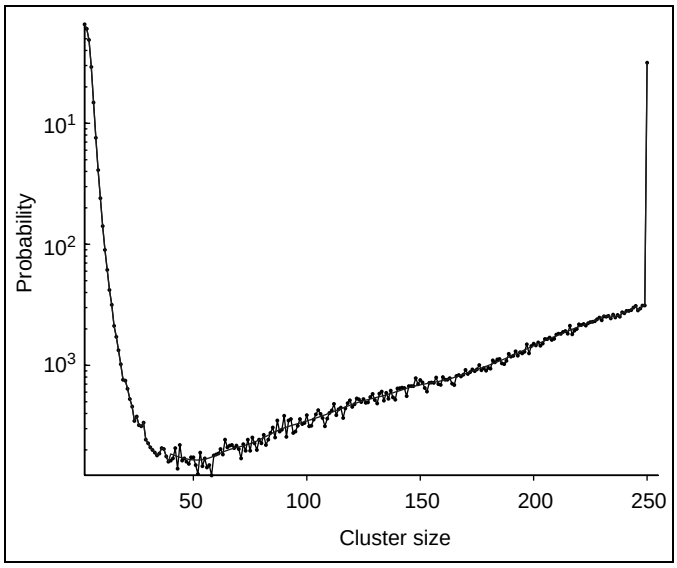

Figure 3. The probability of observing a paralogon by chance can actually increase with the size of the paralogon. The probability of observing at least one cluster of size *h* by chance, when conducting a comparison of two maps each containing 1000 genes and 250 homologous gene pairs, when *d*, the number of interlopers permitted between adjacent paralogs, is 20. The probabilities were estimated from 100 000 randomly permuted genomes.

Concluding remarks

The rapid growth in whole-genome sequencing and the availability of genomic sequence from closely related organisms has launched several studies on large-scale gene duplication. The beginnings of a formal methodology for diagnosing duplications are emerging from this research.

The past five years have seen the birth of generative models of gene duplication times, building on a foundation of established methods for estimating evolutionary divergence of individual homologous pairs. In these studies, each alternate hypothesis is modeled as a combination of small- and large-scale duplication processes. Alternate hypotheses are tested by comparing the predicted outcome of each model with the observed data, although the use of formal statistical tests in comparing predicted and observed outcomes is still emerging.

The use of generative models is a major step towards quantitative comparison of alternative hypotheses, but these models should be used judiciously. A fundamental limitation of the quantitative approach is that several hypotheses can result in the same pattern and, hence, cannot be distinguished using temporal analysis alone [25]. Moreover, quantitative approaches can give a false sense of security. Simplifying assumptions and lack of rigor in comparing observations with models can lead to unfounded conclusions.

The range of results based on spatial data reflects the wealth of different types of information that spatial analysis can offer. These approaches compare the spatial organization of paralogs with the spatial patterns thought to be characteristic of various duplication processes. The use of statistics in these studies is increasing, although statistical testing is often based on randomization without formal statements of the hypotheses and test statistics that are being considered. Typically, putative duplicated regions are tested against null hypotheses of random gene order, rather than more-biologically motivated null hypotheses. We do not have sufficient knowledge about the rates and sizes of large-scale duplications or the rearrangements that fragment them to construct generative models of the sort used in temporal analyses.

Box 4. Future directions

New methodology

Formal approaches for analyzing spatial genomic data are still in their infancy, reflecting the complexity of the problem and our lack of knowledge of the underlying evolutionary processes that shape this data. Although paralogon definitions are often based on hypotheses about how genomic rearrangements proceed, little is known about the rates at which these evolutionary processes occur. The little that is known is often based, circularly, on inferred homologous chromosomal segments. Most algorithmic work for reconstructing genomic rearrangements [75] is parsimony-based, and assumes a small set of rearrangement events that rarely include duplications. The availability of genomic sequences for many closely related species will enable the use of generative models and other statistical formalisms to model rearrangement of large-scale duplicated regions. In the short term, however, there is likely to be little consensus on desirable properties for paralogon definitions.

Combining space and time

Formal approaches that combine temporal and spatial data in a single model are an obvious direction for future work, particularly for increasing the reliability of diagnosis of individual duplication mechanisms. The limitations and sources of error associated with temporal analyses are often orthogonal to those of spatial analyses, whereas their strengths are complementary. Spatial information can be used to tease apart distinct, simultaneous processes that are difficult to distinguish by temporal analysis, whereas similar processes that are difficult to distinguish by spatial analysis can be disambiguated using temporal information if they occurred at different times.

More-powerful data sets

New genomic data will lead to improved models of the rates and characteristics of duplications, and the processes of rearrangement that make ancient large-scale duplications difficult to detect. In addition, studies of recent, naturally occurring polyploids and laboratory generated polyploids are increasing our understanding of the rapid changes that follow aneuploidy and polyploidy [34,76–78]. The results of such studies are challenging accepted views of how and when gene duplications occur and why they persist [14,19,35–37,79]. These new data are revealing substantial, lineage specific differences in the frequency and type of duplications that occur. Lineage specific variation in the distribution of elements that contribute to gene duplication and functional innovation is a challenge for comparative genomics research. Novel approaches will be required to determine to what extent lessons learned from one genome are relevant to another.

Currently, little can be said quantitatively about how well the data supports the various alternate hypotheses under consideration.

Most of the existing methods are focused on global questions concerning the processes that contributed to the evolution of a particular genome; for example, did two polyploidizations occur in early vertebrate evolution [65]? By contrast, local questions focus on the history of a specific gene or region; for example, did the major histocompatibility region (MHC) region arise through an early polyploidization [66]? Because temporal analyses model the duplication-time distribution of a population of paralogs, this approach is used solely to address global questions. In spatial analyses, although the focus is typically on global questions, the duplication mechanism that gave rise to a single pair of paralogs can sometimes be inferred (Box 4). However, it is important to recognize that statistical validation of a paralogon does not constitute evidence for the mechanism of origin of any particular gene in that paralogon. Special care should be taken to avoid interpreting strong support for the global hypothesis of polyploidization as evidence that any individual gene, or even paralogon, arose through WGD.

Acknowledgements

D.D. was supported by NIH grant 1 K22 HG 02451–01 and a David and Lucille Packard Foundation fellowship. R.H. was supported by a Barbara Lazarus Women@IT Fellowship and by the Alfred P. Sloan Foundation. We thank David Sankoff, Robbie Sedgewick, George Weiblen and Ken Wolfe for helpful discussions, and Guillaume Blanc and Ken Wolfe for providing the data used in Figure I.

References

1 *Arabidopsis* Genome Initiative. (2000) Analysis of the genome sequence of the powering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
2 Vision, T.J. *et al*. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* 290, 2114–2117
3 McLysaght, A. *et al*. (2002) Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31, 200–204
4 Panopoulou, G. *et al*. (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.* 13, 1056–1066
5 Long, M. *et al*. (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875
6 Jurka, J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.* 8, 333–337
7 Van de Peer, Y. (2004) Computational approaches to unveiling ancient genome duplications. *Nat. Rev. Genet.* 5, 752–763
8 Hurles, M. (2004) Gene duplication: the genomic trade in spare parts. *PLoS Biol.* DOI:10.1371/journal.pbio.0020206 (http://biology.plosjournals.org)
9 Seoighe, C. (2003) Turning the clock back on ancient genome duplication. *Curr. Opin. Genet. Dev.* 13, 636–643
10 Wolfe, K.H. (2001) Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* 2, 33–41
11 Davis, J.C. and Petrov, D.A. (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* 21, 548–551
12 Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16, 1679–1691
13 Maere, S. *et al*. (2005) Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5454–5459
14 Koszul, R. *et al*. (2004) Eukaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J.* 23, 234–243
15 Patau, K. (1964) Partial trisomy. In *Second International Congress on Congenital Malformations* (Fishbein, M., ed.), pp. 52–59, International Medical Congress
16 Buzdin, A.A. (2004) Retroelements and formation of chimeric retrogenes. *Cell. Mol. Life Sci.* 61, 2046–2059
17 Li, W.H. (1997) *Molecular Evolution*, Sinauer Associates Inc.
18 Zdobnov, E.M. *et al*. (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res.* 33, 946–954
19 Samonte, R.V. and Eichler, E.E. (2002) Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* 3, 65–72
20 Linardopoulou, E.V. *et al*. (2005) Human subtelomeres are hotspots of interchromosomal recombination and segmental duplication. *Nature* 437, 94–100
21 Bailey, J.A. *et al*. (2003) An *Alu* transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* 73, 823–834
22 Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* 291, 1304–1351
23 Kazazian, H.H. (2004) Mobile elements: drivers of genome evolution. *Science* 303, 1626–1632
24 Dujon, B. *et al*. (2004) Genome evolution in yeasts. *Nature* 430, 35–44
25 Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16, 1667–1678

26 Blakeslee, A.F. *et al*. (1920) Chromosomal duplication and mendelian phenomena in datura mutants. *Science* 52, 388–390
27 Hughes, T.R. *et al*. (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.* 25, 333–337
28 Panopoulou, G. and Poustka, A.J. (2005) Timing and mechanism of ancient vertebrate genome duplications. The adventure of a hypothesis. *Trends Genet.* 21, 559–567
29 Jaillon, O. *et al*. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431, 946–957
30 Meyer, A. and Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *BioEssays* 27, 937–945
31 Venkatesh, B. (2003) Evolution and diversity of fish genomes. *Curr. Opin. Genet. Dev.* 13, 588–592
32 Gianni, L. and Louis, E.J. (2005) Yeast Evolution and Comparative Genomics. *Annu. Rev. Microbiol.* 59, 135–153
33 Wolfe, K. (2004) Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol.* 14, R392–394
34 Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141
35 Otto, S.P. and Whitton, J. (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.* 34, 401–437
36 Mable, B.K. (2003) Breaking down taxonomic barriers in polyploidy research. *Trends Plant Sci.* 8, 582–590
37 Mable, B.K. (2004) Why polyploidy is rarer in animals than in plants: myths and mechanisms. *Biological Journal of the Linnean Society* 82, 453–466
38 Soltis, P.S. and Soltis, D.E. (2000) The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7051–7057
39 Bromham, L. and Penny, D. (2003) The modern molecular clock. *Nat. Rev. Genet.* 4, 216–224
40 Gu, X. *et al*. (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat. Genet.* 31, 205–209
41 Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155
42 Lynch, M. and Conery, J.S. (2003) The evolutionary demography of duplicate genes. *J. Struct. Funct. Genomics* 3, 35–44
43 Cotton, J.A. and Page, R.D.M. (2005) Rates and patterns of gene duplication and loss in the human genome. *Proc. Biol. Sci.* 272, 277–283
44 Zhang, P. *et al*. (2004) Different age distribution patterns of human, nematode, and *Arabidopsis* duplicate genes. *Gene* 342, 263–268
45 Wang, X. *et al*. (2005) Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* 165, 937–946
46 Paterson, A.H. *et al*. (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9903–9908
47 The *C. elegans* Sequencing Consortium. (1998) Genome sequence of the nematode *Caenorhabditis elegans*. A platform for investigating biology. *Science* 282, 2012–2018
48 Goff, S.A. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). *Science* 296, 92–100
49 Simillion, C. *et al*. (2004) Recent developments in computational approaches for uncovering genomic homology. *BioEssays* 26, 1225–1235
50 Dietrich, F.S. *et al*. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304, 304–307
51 Kellis, M. *et al*. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624
52 Vandepoele, K. *et al*. (2002) Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.* 18, 606–608
53 Wong, S. *et al*. (2002) Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. U. S. A.* 99, 9272–9277
54 Calabrese, P.P. *et al*. (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* 19(Suppl. 1), i74–i80
55 Durand, D. and Sankoff, D. (2003) Tests for gene clustering. *J. Comput. Biol.* 10, 453–482
56 He, X. and Goldwasser, M.H. (2005) Identifying conserved gene clusters in the presence of homology families. *J. Comput. Biol.* 12, 638–656
57 Hoberman, R. *et al*. (2005) The statistical analysis of spatially clustered genes under the maximum gap criterion. *J. Comput. Biol.* 12, 1083–1102
58 Raghupathy, N. and Durand, D. (2005) Individual gene cluster statistics in noisy maps. In *RECOMB Workshop on Comparative Genomics* (Vol. 3678 of Lecture Notes in Bioinformatics) (McLysaght A. and Huson, D.H. eds), pp. 106–120, Springer-Verlag
59 Vandepoele, K. *et al*. (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15, 2192–2202
60 Simillion, C. *et al*. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* 14, 1095–1106
61 Guyot, R. and Keller, B. (2004) Ancestral genome duplication in rice. *Genome* 47, 610–614
62 Yu, J. *et al*. (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* DOI: 10.1371/journal.pbio.0030038 (http://biology.plosjournals.org)
63 Hoberman, R. and Durand, D. (2005) The incompatible desiderata of gene cluster properties. In *RECOMB Workshop on Comparative Genomics* (Vol. 3678 of Lecture Notes in Bioinformatics) (McLysaght, A. and Huson, D.H. eds), pp. 73–87, Springer-Verlag
64 Bergeron, A. *et al*. (2002) The algorithmic of gene teams. In *WABI* (Vol. 2452 of Lecture Notes in Computer Science) (D. Gusfield and R. Guigo, eds), pp. 464–476
65 Ohno, S. (1970) *Evolution by genome duplication*, Springer-Verlag
66 Danchin, E.G.J. *et al*. (2003) Conservation of the MHC-like region throughout evolution. *Immunogenetics* 55, 141–148
67 Bailey, J.A. *et al*. (2004) Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* 14, 789–801
68 Ciccarelli, F.D. *et al*. (2005) Complex genomic rearrangements lead to novel primate gene function. *Genome Res.* 15, 343–351
69 Tuzun, E. *et al*. (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* 14, 493–506
70 Zhang, L. *et al*. (2005) Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.* 22, 135–141
71 Horvath, J.E. *et al*. (2005) Punctuated duplication seeding events during the evolution of human chromosome 2p11. *Genome Res.* 15, 914–927
72 Cheng, Z. *et al*. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437, 88–93
73 Zhou, Y. and Mishra, B. (2005) Quantifying the mechanisms for segmental duplications in mammaliam genomes by statistical analysis and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4051–4056
74 Long, M. and Thornton, K. (2001) Gene duplication and evolution. *Science* 293, 1551
75 Pevzner, P.A. (2000) *Computational Molecular Biology: An Algorithmic Approach*, MIT Press
76 Adams, K.L. and Wendel, J.F. (2005) Novel patterns of gene expression in polyploid plants. *Trends Genet.* 21, 539–543
77 Andalis, A.A. *et al*. (2004) Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*. *Genetics* 167, 1109–1121
78 Henry, I.M. *et al*. (2005) Aneuploidy and genetic variation in the *Arabidopsis thaliana* triploid response. *Genetics* 170, 1979–1988
79 Dunham, M.J. *et al*. (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16144–16149