Gene and genome duplication

David Sankoff

Genomic sequencing projects have revealed the productivity of processes duplicating genes or entire chromosome segments. Substantial proportions of the yeast, *Arabidopsis* and human gene complements are made up of duplicates. This has prompted much interest in the processes of duplication, functional divergence and loss of genes, has renewed the debate on whether an early vertebrate genome was tetraploid, and has inspired mathematical models and algorithms in computational biology.

Addresses
Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montreal, Québec H3C 3J7, Canada; e-mail: sankoff@poste.umontreal.ca

Current Opinion in Genetics & Development 2001, **11**:681–684

0959-437X/01/$ – see front matter
© 2001 Elsevier Science Ltd. All rights reserved.

Abbreviation
Myr million years

Introduction

There are a number of different ways in which duplicate genes can arise: tandem repeat through slippage during recombination, gene conversion, horizontal transfer and other transposition, hybridization, duplication of entire segments of chromosomes and temporary (auto- or allo-) tetraploidy leading, through processes of diploidization, to an effective doubling of the whole genome. Whatever their origin, duplicate genes may have three kinds of fate. Both copies may persist in the genome with perfect (or near-perfect) sequence identity, possibly resulting in a higher level of expression of the gene product. Alternatively one copy is suppressed, either by physical deletion or by accumulating point mutations until it becomes a pseudogene. Finally, random mutations may cause at least one of the two copies to diverge functionally, either by finding a novel functional role, or by specializing some aspect of its previous role. Iterations and combinations of these processes can give rise to gene families containing from two to hundreds of more or less similar genes carrying out similar or divergent functions.

Several international workshops have focused on this topic recently, including one on "Gene Order Dynamics, Comparative Maps and Multigene Families" organized by myself and JH Nadeau [1•] in Ste-Adèle, Canada in September 2000, another on "Whole Genome Analysis", organized by D Durand at Rutgers University, USA, in February 2001, and a third in Aussois, France in April 2001, organized by A Meyer and H Phillipe. In this review, I concentrate on new information about the prevalence of duplications in various genomes, rates of duplication and loss, which mechanisms are responsible for the observed patterns of duplication with respect to chromosomal position, and the controversy about the role of temporary tetraploidization. I also summarize some mathematical modeling and algorithmics inspired by duplication phenomena.

Gene duplication

Li *et al.* [2] find that duplicated genes, as identified through fairly selective criteria, account for ~15% of the protein genes in the human genome (counting both genes in each pair). In a survey of eukaryotic genome sequences, Lynch and Conery [3••], using a somewhat different filter, accounted for ~8%, 10% and 20% of the gene complement of the fly, yeast and worm genomes, respectively. (Other estimates put the figure at 16% for yeast and 25% for *Arabidopsis* [4•].) They estimated highly variable rates of gene duplication, averaging ~0.01 per gene per Myr (million years). On the basis of ratios of silent and replacement rearrangements, they found that there is typically a period of neutral or (occasionally) even slightly accelerated evolution, lasting a few Myr at most, with one of the copies eventually being silenced in a large majority of cases, and the remaining ones undergoing relatively stringent purifying selection. In contrast, many prokaryotes are susceptible to higher rates of duplication, but with a much more rapid onset of duplicate-gene eradication mechanisms. These authors, along with Force [5–7] and a number of others, make much of 'subfunctionalization', whereby the functional novelty acquired by one or both of the duplicates consists of a specialization of its activity to particular developmental periods, particular tissues, and so on, losing the generality of the ancestral gene. Though supported by many examples, these speculations seem logically independent of the systematic results on duplication and duplication-loss rates provided in this study. Recent work on yeast mutations [8,9] has shown recently that gene duplication does not play a major role in the redundancy of genetic networks, both copies of a duplicate pair tending to be equally essential, with unique functionality.

Segment duplication

Prior to genome sequencing, analyses of duplicated genes could not usually take into account the totality of the chromosomal environment of the genes. It is now clear that many duplicated genes are part of larger duplicated segments. Several studies have explored such segments at two scales of magnitude. Some focused on recent (i.e. >90% sequence similarity, most often >95%) duplications of segments of size mostly in the range of 10–50 kb, though some are only 1 kb and others may be 200 kb. Other studies searched for traces of ancient duplications where the set of genes involved may span many megabases and where successive pairs of matched genes are not necessarily contiguous in either genome, and may indeed be separated by long stretches of unrelated sequence, including many other unduplicated genes.

O'Keefe and Eichler have summarized two patterns of recent segmental duplication widespread in the human genome [10••]. One involves chromosome-specific repeats

such as an 18 kb segment (CH16LAR) that recurs 15 times scattered along the length of the p arm of chromosome 16. The other involves interchromosomal duplications where material located on some chromosomal arm is copied to the pericentromeric or subtelomeric regions of one or more other chromosomes, such as a 10 kb segment (ALD) copied from Xq28 to sites near the centromeres of chromosomes 2, 10, 16 and 22, or a 22 kb segment containing olfactory genes that appears near the telomeres of 10 different chromosomes. It is estimated that 5% of the human genome consists of highly conserved repeats of this kind [11]. Not all the repeated segments contain genes or parts of genes, and for those that do, it is not yet known to what extent they are expressed, if at all. Although some of the sequence similarity occurs as a result of conversion processes, comparative primate genomics confirms the origin of the pericentromeric repeats within the hominoid lineage, within the last 12 Myr.

In the context of the Celera project, a search for ancient duplicated segments in the human genome was based on finding three or more pairs of paralogous genes in relative close proximity on two different chromosomes [12]. More than 1000 such 'blocks' were found, most of them containing five or more genes. Piecing these together led to the identification of a number of very long chromosomal segments that may be relics of ancient duplication events — for example, a region containing 33 proteins genes spanning 20 Mb (and 97 protein genes) on chromosome 2 and 63 Mb (and 332 genes) on chromosome 14. The identification of these segments leads to speculation about their relative timing and the hypothesis of whole genome duplication several hundred Myr ago.

Horizontal gene transfer

Lateral transfer of genes from one organism to another is a mechanism for introducing new genes into a genome. However, it may also create paralogs within the host genome if it already contains a homolog of the transferred gene(s). Eisen [13] has published a comprehensive review of horizontal gene transfer.

Tandem duplication

Models to account for non-contiguous duplicated genes have been explored in genomes as different as mitochondria and yeast [14–16]. The recurrent idea is that tandem duplication of chromosomal segments, by well-understood mechanisms of unequal recombination, are followed by episodes of local chromosomal rearrangement and silencing of most of the duplicate genes in one or the other of the duplicate segments. There are, however, other processes of gene or segment transposition likely to be of greater importance [10••].

Genome duplication

Since the proposals of Ohno in the 1960s, the question of whether vertebrates are the product of two rounds of whole-genome duplication has been hotly debated, and this has been intensified in the past year or two. The establishment of a rigorous protocol for proving a doubling event in the case of yeast [17] (but see [15]) has provided solid criteria for assessing whether the pattern of segmental duplications in a genome is evidence either for or against a history that includes tetraploidization. This produced a clear answer in the case of the *Arabidopsis* genome but not yet for the human genome.

Wolfe [4•] has called for more investigation of the processes of diploidization, whereby a tetraploid, characterized by quadrivalent meiotic figures is returned to a normal state of diploidy. This clearly does not happen instantaneously, even on the evolutionary time scale, and evidence for successive chromosome by chromosome diploidization from the allotetraploid state is available from maize [18,19] and from the autotetraploid state in Salmonid fishes, where the transition from tetraploidy to diploidy is still incomplete and multivalent figures are observed [20].

Arabidopsis genome

Though it has been clear for some time that the *Arabidopsis thaliana* genome has undergone a great deal of segmental duplication [21–26], a convincing analysis has now been published showing that the entire genome was duplicated ~112 Myr ago [27], with no evidence (such as triplicated segments) of multiple independent segmental duplication events or multiple episodes of whole-genome duplication. This is also supported by an independent estimate of duplication times [3••] which shows a relatively sharp peak — though at a somewhat more recent time — as well as comparative mapping evidence [28]. Other authors reconstruct a much more complicated history [29], but this is likely an artifact of inappropriate phylogenetic methodology [4•].

Human genome

The publication of the human genome sequence early in 2001 did not by itself help resolve the long-standing controversy of whether the vertebrate lineage leading to the jawed fishes and thence eventually to humans underwent two rounds of genome duplication. Both the paper from the Celera project [12] and that from the Human Genome Sequencing Consortium [11] mention the extensive segmental duplication that can be detected in the genome but declare that current analyses are insufficient to determine whether these are the result of two whole-genome doublings or a larger number of unrelated duplications of chromosomal segments. Several authors have adduced new evidence in favor [30–33] of the doubling hypothesis and against it [34]. The extreme position against it is advanced by Hughes [35–37] who has marshaled several lines of evidence, most notably a type of phylogenetic analysis in which the four duplicate genes descending from a single ancestor after two genome-doubling events should group together as (AB)(CD) and not as A(B(CD)). The former pattern does not show up more often than would be expected by chance. Both human genome publications reflect the stance taken by Wolfe and co-workers [4•,38] on the basis of a critical assessment of the available data for and against the genome-doubling hypothesis, including an evenhanded assessment

of Hughes' argumentation. Most recently, McLysaght, Hokamp and Wolfe (personal communication) have extracted evidence for a highly significant prevalence of segmental duplications in a relatively short time frame ~450 Myr ago, suggestive of at least one round of tetraploidization.

Algorithms and models

Comparative genomics has given rise to a computational methodology based on gene order rather different from techniques for comparing either nucleotide or amino acid sequences. Here, instead of nucleotide replacements, insertions and deletions, the mathematical comparisons are based on chromosomal rearrangements, such as inversions and translocations [39]. These methods, some requiring rather elaborate theoretical development, are predicated on the hypothesis that the gene orders of two genomes being compared are basically permutations of each other, which is fine for certain small genome comparisons (e.g. metazoan mitochondrial genomes). With larger genomes, most especially the higher eukaryotes, this approach runs into the problem of duplicate genes, paralogy and gene families in general. The gene order of one genome is no longer a permutation of another, and requires a more general type of mathematical description, so that existing algorithms cannot be applied. A solution is found for this generalized version of the gene order comparison problem, where each gene may be present in a number of copies in the same genome, in the notion of 'exemplars', single members of each gene family in each of the two genomes [40]. Exemplars are all identified simultaneously so as to minimize gene order differences between the genomes when all non-exemplars are deleted, so that existing, permutation-based, algorithms are again applicable.

Exemplar analysis may be justified in terms of the biology of rearrangement and duplication processes but is most clearly relevant in the phylogenetic context, where the object is to reconstruct ancestral gene orders on the basis of a given phylogeny. Permutation-based algorithms are available for simple genomes, and are even applicable to the case where some genes are absent from some genomes, but not where genomes may contain gene families. To handle the latter case, we need the additional data represented by the 'gene trees' for each gene family in the data genomes, as produced by standard phylogenetic programs for comparing nucleotide sequence data. Reconciliation analysis [41,42] for projecting the gene trees on the given phylogeny (the 'species tree') can then be used to reconstruct the gene content of the ancestral nodes as well as the ancestral lineages for each member of each gene family. Then, to reconstruct gene orders, we combine exemplar analysis with any of the existing permutation-based gene order phylogeny programs [43•]. Here, we consider, for each ancestral genome (including the root of the tree) and any of its immediate descendants, for each gene in the ancestor, all of its copies produced by duplication processes in the descendant, to be a 'family' for the purposes of the exemplar analysis.

In another algorithmic development, the problem of reconstructing the gene order of the chromosomes of a genome just as it underwent tetraploidization, based on the rearranged (translocated and inverted) genome of a modern-day descendant, has been solved in complete generality by El-Mabrouk [44•]. Her complex but rapid algorithm requires only the chromosomal gene orders and knowledge of all pairs of paralogs resulting from the genome doubling.

Turning from the algorithmics to modeling, a number of abstract probabilistic models have been produced for the generation of multigene families [45–47]. Though of theoretical interest, for the moment they seem little connected to known dynamics of gene duplication.

Acknowledgements

This work supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada. The author is a Fellow of the Program in Evolutionary Biology of the Canadian Institute for Advanced Research.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Sankoff D, Nadeau JH (Eds): **Comparative genomics: gene order dynamics, map alignment and the evolution of gene families, vol 1.** In *Series in Computational Biology.* Dordrecht, NL: Kluwer Academic Press; 2000.

A collection surveying comparative maps and rearrangements from the genetic, cytogenetic, molecular, statistical and algorithmic points of view. Includes studies of organelle, prokaryotic and nuclear genomes and many studies that focus on the implications of gene duplication for genomic evolution

2. Li WH, Gu Z, Wang H, Nekrutenko A: **Evolutionary analyses of the human genome.** *Nature* 2001, **409**:847-849.

3. •• Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290**:1151-1155.

An original and carefully conceived method based on proportion of silent mutations, to trace the trajectory of all detectable pairs of duplicate genes in eukaryotic genomes. The aggregate results show a clear pattern starting with near-neutral evolution and progressing with age towards strong purifying selection.

4. • Wolfe KH: **Yesterday's polyploids and the mystery of diploidization.** *Nat Rev Genet* 2001, **2**:333-341.

A balanced evaluation of the present state of data and analysis pertinent to Ohno's hypothesis of two rounds of early vertebrate genome duplication. The author pays particular attention to how phylogenetic attempts to disprove this hypothesis may be invalidated by assumptions that take into account neither the possibility of allotetraploidy instead of autotetraploidy, nor how asynchronous and drawn-out the return to diploidy may be.

5. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154**:459-473.

6. Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J: **Preservation of duplicate genes by complementary, degenerative mutations.** *Genetics* 1999, **151**:1531-1545.

7. Force A: **The probability of duplicate-gene retention by subfunctionalization and neofunctionalization (abstract).** In *Gene and Genome Duplications and the Evolution of Novel Gene Functions.* Aussois, France: CNRS; 2001.

8. Wagner A: **The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions.** *Genetics* 2000, **154**:1389-1401.

9. Wolfe K: **Robustness — it's not where you think it is.** *Nat Genet* 2000, **25**:3-4.

10. •• O'Keefe C, Eichler E: **The pathological consequences and evolutionary implications of recent human genomic duplications.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:29-46.

The authors review recent discoveries of two segmental duplication processes. One results in copies of certain chromosomal regions appearing near the centromeres or telomeres of several other chromosomes. The other involves multiple copies of a segment scattered along the length of a single chromosome.

11. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.

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

13. Eisen JA: **Horizontal gene transfer among microbial genomes: new insights from complete genome analysis.** *Curr Opin Genet Dev* 2000, **10**:606-611.

14. Boore JL: **The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:133-147.

15. Llorente B, Malpertuy A, Neuveglise C, de Montigny J, Aigle M, Artiguenave F, Blandin G, Bolotin-Fukuhara M, Bon E, Brottier P *et al*: **Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*.** *FEBS Lett* 2000, **22**:101-112.

16. Achaz G, Coissac E, Viari A, Netter P: **Analysis of intrachromosomal repeats in yeast *Saccharomyces cerevisiae*: a possible model for their origin.** *Mol Biol Evol* 2000, **17**:1268-1275.

17. Wolfe KH, Shields DC: **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997, **387**:708-713.

18. Gaut BS, Doebly JF: **DNA sequence evidence for the segmental allotetraploid origin of maize.** *Proc Natl Acad Sci USA* 1997, **94**:6809-6814.

19. Gaut BS, Le Thierry d'Ennequin M, Peek AS, Sawkins MC: **Maize as a model for the evolution of plant nuclear genomes.** *Proc Natl Acad Sci USA* 2000, **97**:7008-7015.

20. Sakamoto T, Danzmann RG, Gharbi K, Howard P, Ozaki A, Khoo SK, Woram RA, Okamoto N, Ferguson MM, Holm LE *et al.*: **A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates.** Genetics 2000, **155**:1331-1345.

21. Ku HM, Vision T, Liu J, Tanksley SD: **Comparing sequenced segments of the tomato and *Arabidopsis genomes*: large-scale duplication followed by selective gene loss creates a network of synteny.** *Proc Natl Acad Sci USA* 2000, **97**:9121-9126.

22. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M *et al.*: **Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402**:761-768.

23. Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N *et al.*: **Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*.** *Nature* 1999, **402**:769-777.

24. McLysaght A, Seoighe C, Wolfe KH: **High frequency of inversions during eukaryote gene order evolution.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:47-58.

25. Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang CX, Katsar CS, Lan TH, Lin YR, Ming R, Wright RJ.: **Comparative genomics of plant chromosomes.** *Plant Cell* 2000, **12**:1523-1540.

26. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: **Extensive duplication and reshuffling in the *Arabidopsis* genome.** *Plant Cell* 2000, **12**:1093-1102.

27. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.

28. Grant D, Cregan P, Shoemaker RC: **Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*.** *Proc Natl Acad Sci USA* 2000, **97**:4168-4173.

29. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in *Arabidopsis*.** Science 2000, **290**:2114-2117.

30. Wang Y, Gu X: **Evolutionary patterns of gene families generated in the early stage of vertebrates.** *J Mol Evol* 2000, **51**:88-96.

31. Gibson TJ, Spring J: **Evidence in favour of ancient octaploidy in the vertebrate genome.** *Biochem Soc Trans* 2000, **28**:259-264.

32. Holland PWH: **Gene duplication: past present and future.** *Semin Cell Dev Biol* 1999, **10**:541-547.

33. Holland PWH, Furlong RF, Pollard SL: **Vertebrate genome evolution: homeoboxes, molecular phylogeny and the octoploidy hypotheses. (Abstract.)** In *Gene and Genome Duplications and the Evolution of Novel Gene Functions*. Aussois, France: CNRS; 2001.

34. Martin A: **Is tetralogy true? Lack of support for the 'one-to-four' rule.** *Mol Biol Evol* 2001, **18**:89-93.

35. Hughes AL: *Adaptive Evolution of Genes and Genomes.* New York: Oxford University Press; 1999.

36. Friedman R, Hughes AL: **Gene duplication and the structure of eukaryotic genomes.** *Genome Res* 2001, **11**:373-381.

37. Hughes AL, da Silva J, Friedman R: **Ancient genome duplications did not structure the human Hox-bearing chromosomes.** *Genome Res* 2001, **11**:771-780.

38. Skrabanek L, Wolfe KH: **Eukaryote genome duplication: where's the evidence?** *Curr Opin Genet Dev* 1998, **8**:694-700.

39. Sankoff D, El-Mabrouk N: **Genome rearrangement.** In *Current Topics in Computational Biology.* Edited by Jiang T, Smith T, Xu Y, Zhang M, Xu Y, Zhang M. Cambridge: MIT Press; 2001:in press.

40. Sankoff D: **Genome rearrangements with gene families.** *Bioinformatics* 1999, **15**:909-917.

41. Page RDM, Cotton JA: **Genetree: a tool for exploring gene family evolution.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:525-536.

42. Chen K, Durand D, Farach-Colton M: **Notung: dating gene duplications using gene family trees.** *J Comp Biol* 2000, **7**:429-447.

43. • Sankoff D, El-Mabrouk N: **Duplication, rearrangement and reconciliation.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:537-550

How to do gene order phylogeny when genomes contain multigene families. Integrates algorithms for gene-tree/species-tree reconciliation, exemplar analysis and gene-order phylogeny to produce gene orders for the ancestral genomes of a given phylogenetic tree.

44. • El-Mabrouk N: **Recovery of ancestral tetraploids.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH, Dordrecht, NL: Kluwer Academic Press; 2000:465-477.

An exact solution to the problem of retracing the rearrangements (translocations and inversions) which account for the evolution of a tetraploid to an observed diploid genome with segmental duplications.

45. Tiuryn J, Radomski JP, Slonimski PP: **A formal model of genomic DNA multiplication and amplification.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:525-536.

46. Gu X: **A simple evolutionary model for genome phylogeny based on gene content.** In *Comparative Genomics.* Edited by Sankoff D, Nadeau JH. Dordrecht, NL: Kluwer Academic Press; 2000:515-523.

47. Altenberg L: **Genome growth and the evolution of the genotype-phenotype map.** In *Evolution and Biocomputation.* Edited by Banzhof W, Eeckman FH. Heidelberg: *Springer Verlag Lecture Notes in Computer Science 899*; 1995:205-259.

Other recommended reading

48. Zhang L, Pond SK, Gaut BS: **A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa.** *J Mol Biol* 2001, **52**:144-156.