Double Dutch for duplications

Ewan Birney

A new study presents a comprehensive global analysis of the evolution of segmental duplications in the human genome. The authors identify the origin of ancestral duplication loci, regions of clustered duplicons, and evidence supporting a punctuated model of evolution.

The evolution of genomes involves not only single-base-pair changes or small insertions or deletions, but also duplications and deletions of large portions of the genome. Segmental duplications (or duplicons), segments of DNA with nearly identical sequence, can be hot spots for non-allelic homologous recombination and sites for the occurrence of genomic mutations that include deletions, duplications and inversions[1]. Previously, phylogenetic and comparative genomic studies have examined the evolution of individual segmental duplications, and these studies supported a multistep model for the evolution of these duplications. However, this approach has been limited by the complicated structure of the regions, which made traditional, collinear multiple alignment infeasible. Evan Eichler, Pavel Pevzner and colleagues, reporting on page 1361 of this issue[2], now present an analytical approach that provides the first genome-wide overview of these events. Through an elegant combination of new computational techniques, comparative genomics and targeted experiments, they provide the first comprehensive global analysis of human segmental duplications. They show conclusively that these regions are dynamic and highly variable between species. They also find a core of duplications that account for most of the variability, and show that some of these contained transcribed genes in their copies, strongly suggesting that they have a role in positive selection.

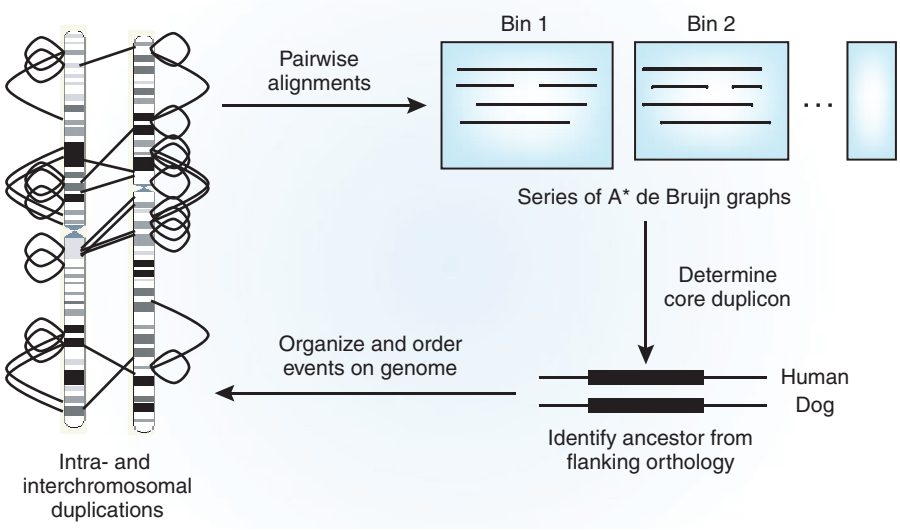

Figure 1 Data flow for resolving segmentation duplications. The input is a set of pairwise alignments, which is then binned into connected sets. Each bin is then subjected to the de Bruijn graph procedure to find a set of underlying duplicons. Using flanking homology, the ancestral duplicon can then be determined. From the set of duplicons and ancestors, the current set of duplicated sequences can be organized and their events ordered in time.

Starting with the reference

Jiang *et al.*[2] draw upon the current dataset based on the reference human genome for segmental duplications in the human genome[1]. This existing dataset consists of over 28,000 pairwise alignments of the reference human genome, which have been carefully calibrated to retain only recent segmental duplication events of reasonable size (as there are a whole series of different transposon insertions or deletions that are not of direct interest to this segmental duplications). From this initial reference dataset, Jiang *et al.* narrowed the set to 11,951 nonredundant duplication subunits. This large dataset has alignments scattered across the genome, many of them in pericentromeric areas with a strong bias toward intrachromosomal duplications. However, this web of pairwise alignments does not directly indicate the ordering of duplications or, critically, which sequence provided the ancestral sequence to the duplication (**Fig. 1**). Resolving these alignments is particularly complex, as not only can one sequence be duplicated multiple times, but also the subsequent duplications themselves can either have regions deleted, be partially duplicated or be completely embedded in larger duplications. The set of pairwise alignments therefore touches the set of extant sequences involved in duplications, and provides a very complex (and oddly beautiful) pictorial representation of segmental duplications, but it does not illuminate the history of these duplications.

De Bruijn graphs to reconstruct ancestral duplications

Jiang *et al.*[2] solved this web of alignments by employing a data structure borrowed from combinatorial mathematics—the de Bruijn graph, created by a Dutch mathematician with

Ewan Birney is at the European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, United Kingdom.
e-mail: birney@ebi.ac.uk

some interests in biology as it related to natural language, but no documented thought that this could be applied to molecular biology. A de Bruijn graph is a compact way to represent strings of letters with a number of useful properties. The critical property utilized here is that any repeated subsequence over a particular length is present only once in the graph, regardless of how many times it may be present in the full sequence. If one imagined the simplified case in which there were only duplication events (and no small base pair changes) giving rise to a set of potentially nested duplications, the de Bruijn graph would have the ancestral duplicated sequence as one set of unique subsequences in the graph. Naturally, biology is not quite as accommodating of this mathematical theory as one would like; however, Pavel Pevzner and colleagues have previously provided manipulations of the de Bruijn graph to effectively 'smooth out' small differences and thus obtain a clean set of duplication subunits which one can then trace to all their subsequent copies[3]. In this regard, Jiang *et al.* offer yet another application of de Bruijn graphs, which are of growing importance for DNA sequence analysis.

As well as using the human input sequence, Jiang *et al.* took a comparative genomics approach drawing on other mammalian genomes. They were immediately able to find the potential ancestral locus using the fact that the ancestral region would be embedded in a longer region of homology to outgroup mammalian species. They predicted around 5,000 such putative ancestral duplication loci, generating over 100 Mb of sequence in the reference human genome. Using a targeted FISH approach, they showed that these predictions of ancestral loci were correct most of the time, with the other cases likely to reflect more complex duplications occurring during primate evolution.

Now armed with a sensible ordering and organization of these duplication events, Jiang *et al.* have provided a far richer descriptive analysis of segmental duplications in the human genome than had previously been possible. As suspected, there are radical differences in the rates of intrachromosomal duplications, with chromosome segments 1q, 8, 9p, 10q, 15q 16p, 17, 19, 22q, X and Yq showing higher rates, but with these same chromosomal regions harboring high levels of both ancestral and derived duplicons, suggesting that this chromosomal bias is in both the generation and the integration of duplications. Many duplications lie in complex blocks of nested duplications, and there is evidence for a burst of duplication activity at a specific point in evolution. Given the association of segmental duplications to positively selected genes in the human genome and other genomes, such a coordinated burst of duplication events suggests a tantalizing hypothesis of a sudden, punctuated set of phenotypic changes in a short period of time—the punctuated model of evolution. This analysis does not show that this burst in duplication is associated with phenotypic changes, but this will be a fascinating hypothesis to explore in the future.

In addition, these complex blocks themselves have inter-relationships, with 24 groups of duplication blocks (the groups often, but not always, concentrated in a particular chromosome). As noted in previous studies, some of these segmental duplications show strong evidence of recent innovation (such as TRE2 gene on chromosome 17), and other duplications show gene families with unknown functions.

Further structural variation analyses

Where does this analysis leave us? Duplications that are fixed in the population, so-called segmental duplications, were the main focus of this study, but this framework is just as important in understanding polymorphic duplications, or copy number variations (CNVs). It has long been suggested that positive selection may have a complex relationship with disease association[4], and the linking of both of these with the common large-scale mutational process of genome duplication provides a new avenue to explore in investigating disease association. Currently, many groups are actively pursuing studies to associate CNVs with disease[5]. These CNVs may have far greater spontaneous or recurrent behavior than the more 'traditional' base pair changes, and may thus change our understanding of the importance of recent as compared to ancestral mutations in the genetics of common disease[6]. Jiang *et al.* now provide a framework to analyze both polymorphic (CNV) and fixed (segmental) duplications, and the relationship between recent positive selection and disease association will be greatly increased by having this catalog and analysis of recent duplication events. Studies such as this one will become increasingly important with the advent of multiple individual genome sequences over the coming years[7].

1. She, X. *et al. Nature* **431**, 927–930 (2004).
2. Jiang, Z. *et al. Nat. Genet.* **39**, 1361–1368 (2007).
3. Pevzner, P.A., Tang, H. & Tesler, G. *Genome Res.* **14**, 1786–1796 (2004).
4. Olson, M.V. & Varki, A. *Nat. Rev. Genet.* **4**, 20–28 (2003).
5. McCarroll, S.A. & Altshuler, D.M. *Nat. Genet.* **39** (Suppl.), S37–S42 (2007).
6. Conrad, D.F. & Hurles, M.E. *Nat. Genet.* **39** (Suppl.), S30–S36 (2007).
7. Levy, S. *et al. PLoS Biol.* **5**, e254 (2007).

How the dog got its spots

Gregory S Barsh

Differences among dog breeds provide unique opportunities for studying the genetics of behavior, morphology and complex disease. Two new studies demonstrate how the unique evolutionary history of domestic dogs is particularly well suited to analysis by genome-wide association.

On May 7 the Rockefeller Foundation announced a grant of $282,000 to the Roscoe B. Jackson Memorial Laboratory for studies over a five-year period of genetic factors in intelligence and emotional variation in mammals. This most interesting project may mark a turning-point in genetic research as it applies to questions of peculiar importance in human society[1].

Gregory S. Barsh is in the Departments of Genetics and Pediatrics, Stanford University School of Medicine, Stanford, California 94305, USA.
e-mail: gbarsh@stanford.edu

One expects to see similar announcements today (hopefully with larger sums of money), but this research project was funded more than 60 years ago, focusing on variation in behavior and morphology among five breeds of domestic dogs: the basenji, beagle, cocker spaniel, Shetland sheep dog and wire-haired fox terrier.

As beautiful examples of the power of selective breeding, differences among dog breeds have intrigued geneticists for nearly a century[2]. Only in the last few years, though, have molecular