

References

- 1 DiMauro, S. and Schon, E.A. (2001) Mitochondrial DNA mutations in human disease. *Am. J. Med. Genet.* 106, 18–26
- 2 Chinnery, P.F. *et al.* (2000) Epidemiology of pathogenic mitochondrial DNA mutations. *Ann. Neurol.* 48, 188–193
- 3 Chinnery, P.F. and Turnbull, D.M. (1999) Mitochondrial DNA and disease. *Lancet* 354 (Suppl. I), 17–21
- 4 Leonard, J.V. and Schapira, A.V.H. (2000) Mitochondrial respiratory chain disorders I: mitochondrial DNA defects. *Lancet* 355, 299–304
- 5 Battersby, B.J. and Shoubridge, E. (2001) Selection of a mtDNA sequence variant in hepatocytes of heteroplasmic mice is not due to difference in respiratory chain function or efficiency of replication. *Hum. Mol. Genet.* 10, 2469–2479
- 6 Wallace, D.C. (1999) Mitochondrial diseases in mouse and man. *Science* 283, 1482–1488
- 7 Larsson, N.G. *et al.* (1992) Segregation and manifestations of the mtDNA tRNA(Lys) A→G(8344) mutation of myoclonus epilepsy and ragged-red fibers (MERRF) syndrome. *Am. J. Hum. Genet.* 51, 1201–1212
- 8 Macmillan, C. *et al.* (1993) Variable distribution of mutant mitochondrial DNAs (tRNA(Leu[3243])) in tissues of symptomatic relatives with MELAS: the role of mitotic segregation. *Neurology* 43, 1586–1590
- 9 Jenuth, J. *et al.* (1996) Random genetic drift in the female germ line explains the rapid segregation of mammalian mitochondrial DNA. *Nat. Genet.* 14, 146–151
- 10 Meirelles, F. and Smith, L.C. (1997) Mitochondrial genotype segregation in a mouse heteroplasmic lineage produced by embryonic karyoplast transplantation. *Genetics* 145, 445–451
- 11 Takeda, K. *et al.* (2000) Replicative advantage and tissue-specific segregation of RR mitochondrial DNA between C57BL/6 and RR heteroplasmic mice. *Genetics* 155, 777–783
- 12 Chinnery, P.F. *et al.* (2000) The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet.* 16, 500–505
- 13 Brown, D.T. *et al.* (2000) Random genetic drift determines the level of mutant mtDNA in human primary oocytes. *Am. J. Hum. Genet.* 68, 533–536
- 14 Chinnery, P.F. *et al.* (1999) Non-random tissue distribution of mutant mitochondrial DNA. *Am. J. Med. Genet.* 85, 498–501
- 15 Dubeau, F. *et al.* (2000) Oxidative phosphorylation defect in the brains of carriers of the tRNA^{(Leu(UUR))} A3243G mutation in a MELAS pedigree. *Ann. Neurol.* 47, 179–185
- 16 Jenuth, J.P. *et al.* (1997) Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. *Nat. Genet.* 16, 93–95
- 17 Bidooki, S.K. *et al.* (1997) Intracellular mitochondrial triplasmia in a patient with two heteroplasmic base changes. *Am. J. Hum. Genet.* 60, 1430–1438
- 18 Holt, I.J. *et al.* (2000) Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* 100, 515–524
- 19 Turnbull, D.M. and Lightowlers, R.N. (1998) An essential guide to mtDNA maintenance. *Nat. Genet.* 18, 199–200
- 20 Lehtinen, S.K. *et al.* (2000) Genotypic stability, segregation and selection in heteroplasmic human cell lines containing np 3243 mutant mtDNA. *Genetics* 154, 363–380
- 21 Elson, J.L. *et al.* (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.* 68, 802–806
- 22 Coller, H.A. *et al.* (2001) High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.* 28, 147–150
- 23 Blier, P.U. *et al.* (2001) Natural selection and the evolution of mtDNA-encoded peptides: evidence for intergenomic co-adaptation. *Trends Genet.* 17, 400–406

Patrick F. Chinnery

Dept of Neurology, The Medical School,
Newcastle upon Tyne, UK NE2 4HH.
e-mail: P.F.Chinnery@ncl.ac.uk

Genome Analysis

Identifying functional links between genes using conserved chromosomal proximity

Itai Yanai, Joseph C. Mellor and Charles DeLisi

Conservation of proximity of a pair of genes across multiple genomes generally indicates that their functions could be linked. Here, we present a systematic evaluation using 42 complete microbial genomes from 25 phylogenetic groups to test the reliability of this observation in predicting function for genes. We find a relationship between the number of phylogenetic groups in which a gene pair is proximate and the probability that the pair belongs to a common pathway. Our method produces 1586 links between ortholog families substantiated by observed proximity in genomes representing at least three phylogenetic groups. Of the pairs annotated in the KEGG database, 80% are in the same biological pathway in KEGG.

The relationship between the physical order of a set of genes and that of its orthologs in other species becomes increasingly random as the phylogenetic distance between the respective genomes

diverges [1–4]. Nevertheless, the numerous correlations that are found [2,4–8] are informative, because conserved order implies an underlying selective bias and thus perhaps related function. The classic operon [9], where conserved proximity preserves a crucial cellular function, is one example. In a 'selfish operon', proximity between genes is maintained mostly because it facilitates their horizontal co-transfer to another organism [10]. For example, in restriction-modification systems, the gene for a restriction enzyme invariably clusters with the corresponding modification methylase gene along the chromosome and 'their behavior appears to reflect a primarily 'selfish' purpose' [11]. Moreover, Dandekar *et al.* [2] have presented evidence that the products of proximate genes tend to interact physically and there is also support for the notion that proximate genes in yeast have similar expression patterns [12]. These observations lead to the question: given a

reasonable definition of proximity, how frequently does it imply function, or, in slightly different terms, how likely is it that two genes are proximate by chance?

Comparative genomics can provide a clue. If there is positive selection for the proximity of two genes, proximity is also expected between their orthologs. Thus, one approach for detecting functionally coupled neighbors is to search for gene proximity that is conserved across genomes. Overbeek and colleagues [13] pioneered this approach and used it to reconstruct several pathways and to predict the participation of new genes in certain pathways. Kanehisa and colleagues [14] used a complementary approach to identify and align orthologous gene clusters based upon graph analysis algorithms.

Although these analyses are promising, there has been no systematic evaluation of the validity of the hypothesis. Here, we systematically evaluate the probability that proximate genes encode proteins in

the same biological pathway, as a function of the number of genomes in which proximity is conserved. The results indicate that predictions can be made with very high precision, depending on the number of genomes in which we require that proximity be conserved.

Finding proximate genes and constructing profiles of conservation

We define genes as proximate if they are on the same strand and within 300 base pairs [13], or if their respective paralogs are within 300 bp (Fig. 1), with paralogs and orthologs defined according to the Clusters of Orthologous Groups database (COG; Box 1) [15,16]. Because the complete genome sequences available are not evenly distributed phylogenetically (e.g. two strains of *Helicobacter pylori*), we adopt the COG categorization of the 42 microbial genomes into 25 major phylogenetic groups (see supplementary information at <http://fusion.bu.edu/org.doc>). We then construct chromosomal proximity profiles for each pair of genes by noting the phylogenetic groups in which they are proximate. We consider two genes to be linked by chromosomal proximity if their profile indicates that they are proximate in more than a threshold number of phylogenetic groups (see below and Fig. 1). All direct and inferred links have been deposited in Predictome, a database of putative links between proteins (<http://predictome.bu.edu>). The site also contains detailed information on the generation of the links.

A chromosomal proximity profile can be used to derive two types of links between genes of a particular organism. If two genes from a given genome are proximate and their orthologs are also proximate in at least one other genome of a different phylogenetic group, the two can be said to be directly linked. In addition, the conservation of proximity between a pair of genes in several genomes can also hint at a functional relationship in an organism where the genes are present but not proximate. This notion is analogous to that of the domain fusion method, where a multidomain fusion protein in one organism suggests a functional interaction between the domains in a genome where they are encoded by distinct genes [17–19]. Thus, we define an inferred 'chromo link' as relating a pair of genes that are not close but whose orthologs are close in three (see below) or more other phylogenetic groups (Fig. 1).

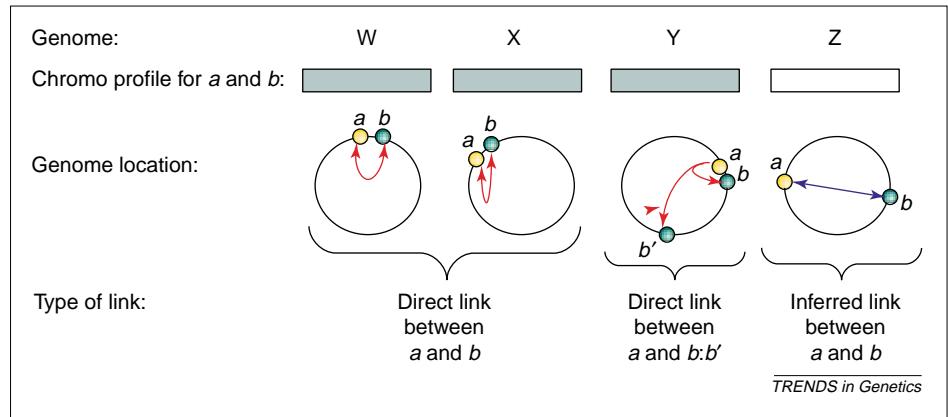


Fig. 1. Direct and inferred chromo links. The four chromosomes are arbitrarily labeled W, X, Y, and Z. The 'chromosome profile', shown above the chromosome diagram, is shaded if genes *a* and *b* are close and left blank otherwise. A direct 'chromo link' can occur between two close genes that are also close in at least two other genomes of different phylogenetic groups (see supplementary information for list of groups). An inferred link relates two genes that are not close but whose orthologs are close in at least three other genomes of different groups. We adopt the Overbeek *et al.* [13] operational definition of 'close': the pair of genes are separated by at most 300 base pairs and transcribed in the same direction. All sets of paralogous genes (for example *b* and *b'*) are collapsed, and so a direct link can relate two genes that are not close (for example, *a* and *b'* in the link between *a* and *b:b'*).

Figure 2 shows the direct and inferred chromo links in *Buchnera* sp. APS of the Enterobacteriaceae group. The direct links are evenly distributed along the chromosome with paralog direct links (see Fig. 1) observed as both tandem to their paralog and separated by distance. Inferred links also show a fairly even spatial distribution and frequently link genes separated by large chromosomal distances. This indicates that some inferred links do not correspond to local rearrangements of the genes but instead a shuffling of the gene order on a chromosomal scale. Had the inferred links corresponded only to relatively proximate genes (i.e. not within 300 bp), the notion of an inferred proximity link would not be warranted. However, because some of the linked genes are distant, the link is inferred solely from the chromosomes of other organisms. From an evolutionary standpoint, this observation suggests that

chromosomal order operates at the local level, for example operons, as opposed to the global order of local structures.

From a practical point of view, these observations indicate that when a genome is mapped, the chromosomal proximity profile between the orthologs of every pair of genes must be obtained, including those that are not proximate. For example, if the set of genes *ribABCDH* in the *Escherichia coli* riboflavin biosynthesis pathway were not known to be related, a relationship would not be suggested merely from a consideration of the local chromosomal organization of *E. coli*, because those genes are widely separated in that organism (with the exception of the proximity between *ribD* and *ribH*). Table 1 shows the links involving the set of *E. coli* genes where at least one of the paired genes is involved in the riboflavin pathway. For example, there is an inferred link between *ribH* and *ribB* because in

Box 1. Databases used for functional correlation analysis

COG: Clusters of Orthologous Groups (<http://www.ncbi.nlm.nih.gov/COG/>)

In this well curated database, orthologous proteins are clustered and assigned a functional category. The 15 broad functional categories range from informational functions, such as 'transcription' and 'DNA replication, recombination, and repair', to metabolic functions, such as 'lipid metabolism' and 'energy production and conversion'. Here, we consider two proteins to be functionally correlated according to the COGs database if they are assigned the same functional category.

KEGG: Kyoto Encyclopedia of Genes and Genomes (<http://www.genome.ad.jp/kegg/>)

In the PATHWAY section of this extensive database, genes are grouped into 138 metabolic and regulatory pathways. Genes are assigned to pathways, such as 'galactose metabolism' and 'MAPK signaling pathway' based on the literature and extended to newly sequenced genomes by sequence similarity. In this work, two genes are deemed functionally correlated if they are both members of the same pathway.

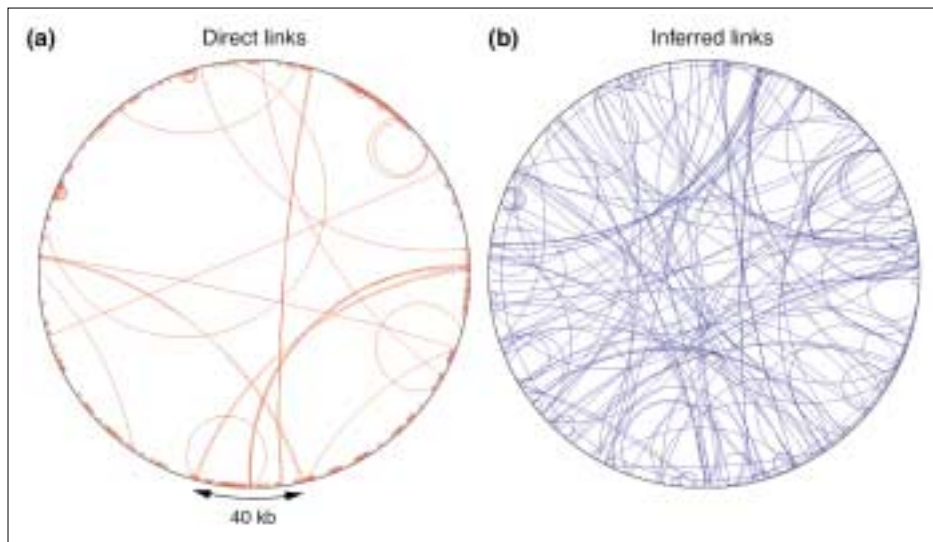


Fig. 2. Visualization of direct (a) and inferred (b) 'chromo links' in the *Buchnera* sp. APS genome. The 231 direct chromo links (red lines) are mostly close on the chromosome, the remaining are due to paralogs (see text and Fig. 1). The 136 inferred links (blue lines) mostly link distant genes. The *Buchnera* sp. APS genome is 640 681 bases.

the 25 phylogenetic groups examined, the two genes are physically proximal (≤ 300 bp) in ten groups: VDRLBFGSJI (see supplementary information for the list of groups). The five inferred links all involve genes known to interact through the riboflavin pathway. Although the genes are not neighbors in *E. coli*, their orthologs in other groups are often proximate.

However, in the two direct links, one gene is in the riboflavin pathway and the other is not. In both of these cases, the second gene is involved in transcription: *nusB* is a transcription antiterminator, and *ybaD* is predicted in the COG database to be a transcription regulator. These genes are adjacent to genes involved in the riboflavin pathway in at least six (*ybaD*) and eight (*nusB*) different groups, indicating that selection is operating to

keep these pairs proximate, presumably because of functional co-dependence. A closer look at *nusB* gene reveals a direct link with *thiL*, the product of which is involved in the biosynthesis of thiamine, another vitamin. As it is thought that *nusB* is involved in a process that is invoked in periods of nutrient deprivation, it is probable that its position between two genes involved in vitamin biosynthesis is not random and involves a selective bias.

How reliable is this method?

To estimate the ability of our method to detect links between functionally related genes, we identify the subset of links where both ortholog families are annotated and assigned to a pathway in the KEGG database [20], or to a functional category in the COG database [15] (Box 1).

We then calculate the fractions of these that are in the same KEGG pathway and of the same COG functional category. As Fig. 3a shows, the functional correspondence is positively correlated to the minimal number of phylogenetic groups in which the proximity is detected (henceforth called N). Thus, specifying a higher N increases the quality of the links. Given a stringent threshold of $N \geq 12$, the resulting links show a functional correlation of 90% or greater against both the KEGG and COG ontologies. However, the pairs show a stronger functional correlation with the KEGG database at all $N \leq 11$. This difference can be attributed to the difference of organizations of the two ontologies (Box 1).

Setting a stringent N threshold has the tradeoff of producing fewer links (Fig. 3b). To arrive at a compromise between the number and quality of links generated, we have set the threshold N to three (Fig. 3). Each link between two ortholog families produces direct links between the genes that are proximate in the respective genomes and inferred links between those genes that are not proximate (Fig. 1). Thus, effectively we define a direct link between two proximate genes in a genome if they are also proximate in at least two other groups. An inferred link relates two non-proximate genes if they are proximate in at least three groups.

We detect 1586 links between ortholog families (COG). Of the links of these where both members are annotated in a KEGG pathway, 80% are present in the same biological pathway. This figure is reduced to 67% correspondence with the COG broad functional categories. Overall, when these links used to produce direct and inferred links between genes, the 42 genomes analyzed average 380 direct and 352 inferred links, yielding totals of 12 755 and 14 352 links, respectively.

A gene can be linked by multiple chromosomal proximity links, and thus collectively the total set of links form a complex network. We intentionally refrain from building clusters of links in this network, as discrete clusters would reduce the amount of information inherent in the network. Instead, investigating an interesting gene requires examination of the linked genes and their linked genes, and so on. An equally important advantage of pairwise links that distinguish between direct and inferred chromosomal proximity is their compatibility of the chromo links

Table 1. Chromo links containing *Escherichia coli* genes involved in the riboflavin biosynthesis pathway

Gene A	COG A	Gene B	COG B	Chromosomal proximity profile ^a	Link
<i>ribH</i>	COG0054	<i>ribB</i>	COG0108	-----VDRLB--FG-S--J-I--	Inferred
<i>ribH</i>	COG0054	<i>ribA</i>	COG0807	-----VDRLB--FG-S--J-I--	Inferred
<i>ribB</i>	COG0108	<i>ribC</i>	COG0307	-----VDRLB--FG-S--J----	Inferred
<i>ribA</i>	COG0807	<i>ribC</i>	COG0307	-----VDRLB--FG-S--J----	Inferred
<i>ribC</i>	COG0307	<i>ribD</i>	COG0117	-----VD-LB--G----J----	Inferred
<i>ribH</i>	COG0054	<i>nusB</i> ^b	COG0781	-----Q-----EFGHSNUJ----	Direct
<i>ybaD</i> ^b	COG1327	<i>ribD</i>	COG0117	-----EFGH-N-J----	Direct

^aEach letter in the profile corresponds to a phylogenetic group as defined in the Cluster of Orthologous Groups database [16] (Box 1). B, the *Bacillus* group; D, the *Deinococcus* group; E, the Enterobacteriaceae, which contains the *E. coli* genome whose genes are considered here; F, the *Pseudomonas* group; G, the *Vibrio* group; H, the Pasteurellaceae; I, the *Chlamydia* group; J, the Proteobacteria α -subdivision; L, the Streptococcaceae; N, the Neisseria; R, the *Mycobacterium* group; S, the *Xylolla* group; U, the ϵ -proteobacteria subdivision; V, the *Thermotoga* group, which includes (only) the *Thermotoga martima* genome. The complete list of the phylogenetic groups is also given in the supplementary information at <http://fusion.bu.edu/org.doc>.

^bGenes not known to be involved in riboflavin biosynthesis.

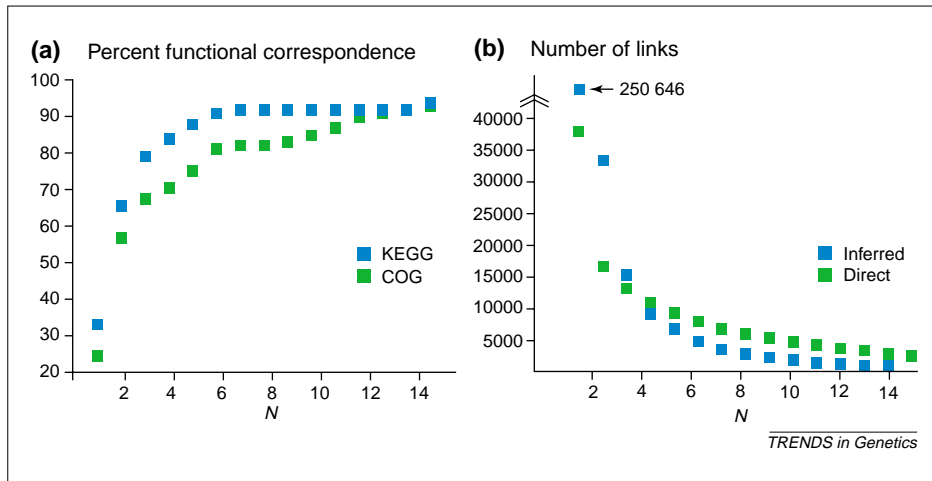


Fig. 3. Functional analysis of chromosomal proximity links (a) Dependence of the functional correspondence on N , the minimum number of phylogenetic groups in which proximity is observed. Functional correspondence in KEGG refers to the participation in the same pathway and in COG to membership in the same functional category. In this study, we define a chromosomal proximity link between two ortholog families (COG) if the proximity between the genes of the families occurs in at least three phylogenetic groups. (b) The dependence of the number of links of the direct and inferred links on N , produced from the links between ortholog families.

with phylogenetic profile links [15,21,22] and fusion links [17–19]. Together, these constitute the links generated through an analysis of the context in which genes are present in genomes.

Acknowledgements

We would like to thank Adnan Derti for a critical reading of this manuscript. This work was supported in part by a National Science Foundation Integrative Graduate Education and Research Traineeship Program grant to JCM. IY is supported by a Whitaker Fellowship.

References

- Mushegian, A.R. and Koonin, E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.* 12, 289–290
- Dandekar, T. *et al.* (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* 23, 324–328
- Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis. *Adv. Protein Chem.* 54, 345–379
- Wolf, Y.I. *et al.* (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* 11, 356–372
- Danchin, A. *et al.* (2000) Mapping the bacterial cell architecture into the chromosome. *Philos. Trans. R. Soc. London Biol. Sci.* 355, 179–190
- Rocha, E.P. *et al.* (2000) Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. *J. Biotechnol.* 78, 209–219
- Suyama, M. and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. *Trends Genet.* 17, 10–13
- Lathe, W.C., III *et al.* (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.* 25, 474–479
- Jacob, F. (1997) The operon after 25 years. *C. R. Acad. Sci. Ser. III* 320, 199–206
- Lawrence, J.G. (1997) Selfish operons and speciation by gene transfer. *Trends Microbiol.* 5, 355–359
- Kobayashi, I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* 29, 3742–3756
- Cohen, B.A. *et al.* (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26, 183–186
- Overbeek, R. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2896–2901
- Fujibuchi, W. *et al.* (2000) Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping. *Nucleic Acids Res.* 28, 4029–4036
- Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
- Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
- Marcotte, E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science* 285, 751–753
- Enright, A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90
- Yanai, I. *et al.* (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7940–7945
- Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30
- Gaasterland, T. and Ragan, M.A. (1998) Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* 3, 177–192
- Pellegrini, M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4285–4288

Itai Yanai*

Joseph C. Mellor

Charles DeLisi

Bioinformatics Graduate Program and Dept of Biomedical Engineering, Boston University, Boston, MA 02215, USA.

*e-mail: iyanai@bu.edu

Meeting Report

Insects on the rise

Diethard Tautz

The Insect Genomics Workshop was held in Arlington, Virginia, from 28 to 30 October 2001.

'*Drosophila* is an insect,' was the opening statement of Mike Asburner's presentation at the workshop. But this is only partly true, because for the genomics initiatives it is a model organism. This is

why it got funded within the human genome consortium. However, it was made clear by Francis Collins (NIH, Bethesda, MD, USA) that this will not happen for other insects. Insect genomics needs to find money from other sources, the most obvious ones being those that are linked to agricultural interests. Damage caused by pest insects result in a loss of at

least 20% of the world harvest each year, amounting to a loss of at least 100 billion dollars. On the other hand, revenues generated from beneficial insects like honey bees or silk moths are worth a similar sum and need increasing protection. The workshop came therefore at the right time and the organizers (Kevin Hackett, Agricultural Research