Accepted papers

Reconstructing Domain Compositions of Ancestral Multi-domain Proteins
Behshad Behzadi and Martin Vingron

Domain Architecture in Homolog Identification
Nan Song, Robert Sedgewick, and Dannie Durand

Inferring positional homologs with common intervals of sequences
Guillaume Blin, Annie Chateau, Cedric Chauve, Yannick Gingras

On genome evolution with accumulated change and innovation
Damian Wojtowicz and Jerzy Tiuryn

Paths and Cycles in Breakpoint Graphs of Random Multichromosomal Genomes
Wei Xu, Chunfang Zheng and David Sankoff

Common Intervals and Symmetric Difference in a Model-free Phylogenomics, with an Application to Streptophyte Evolution
Zaky Adam, Monique Turmel, Claude Lemieux, and David Sankoff

How Pseudo-Boolean Programming can help Genome Rearrangement Distance Computation
Sébastien Angibaud, Guillaume Fertin, Irena Rusu, and Stéphane Vialette

Sorting by translocations via reversals theory
Michal Ozery-Flato and Ron Shamir

Inferring gene orders from gene maps using the breakpoint distance
Guilllaume Blin, Eric Blais, Pierre Guillon, Mathieu Blanchette, and Nadia El-Mabrouk

Ordering Partially Assembled Genomes Using Gene Arrangements
Éric Gaul and Mathieu Blanchette

Evolution of Tandemly Repeated Sequences through Duplication and Inversion
Denis Bertrand, Mathieu Lajoie, Nadia El-Mabrouk, and Olivier Gascuel

A PQ Framework for Reconstructions of Common Ancestors & Phylogeny
Laxmi Parida

Intron Loss Dynamics in Mammals
Jasmin Coulombe-Huntington and Jacek Majewski

Finding Maximum Likelihood Indel Scenarios
Abdoulaye Banir´e Dial lo, Vladimir Makarenkov, and Mathieu Blanchette

Conservation Patterns in Cis-elements Reveal Compensatory Mutations
Perry Evans, Greg Donahue, and Sridhar Hannenhalli

Transcription Factor Centric Discovery of Regulatory Elements in Mammalian Genomes Using Alignment-Independent Conservation Maps
Nilanjana Banerjee and Andrea Califano

Identifiability Issues in Phylogeny-based Detection of Horizontal Gene Transfer
Cuong Than, Derek Ruths, Hideki Innan, and Luay Nakhleh

Invited Talks

Fiona Brinkman
Evolution of microbial cellular networks: Insights from improved genome-wide prediction of orthologs and protein subcellular localization

We have developed the Ortholuge and PSORTb computational methods for more precise prediction of orthologs and protein subcellular localization, respectively. By applying these methods for the comparative analysis of all complete bacterial and archaeal genomes, we are able to identify statistically significant trends in the structure of microbial cellular networks as a function of genome size. Our results support a model whereby subnetworks (that are added to an existing basic network) are primarily responsible for movement of substrates or products in or out of the cell. The implications for the evolution of large prokaryotic genomes is discussed in the context of horizontal gene transfer. Evidence is also presented demonstrating that we must improve our ortholog predictions for comparative analyses, since our models suggest that roughly 1 in 10 eukaryotic, and 1 in 20 bacterial, ortholog pairs are being miss-predicted using the common reciprocal-best-BLAST-hit method for ortholog identification.

Tao Jiang
A Parsimony Approach to Genome-Wide Ortholog Assignment

The assignment of orthologous genes between a pair of genomes is a fundamental and challenging problem in comparative genomics. Existing methods that assign orthologs based on the similarity between DNA or protein sequences may make erroneous assignments when sequence similarity does not clearly delineate the evolutionary relationship among genes of the same families. In this paper, we present a new approach to ortholog assignment that takes into account both sequence similarity and evolutionary events at genome level, where orthologous genes are assumed to correspond to each other in the most parsimonious evolving scenario under genome rearrangement and gene duplication. It is then formulated as a problem of computing the signed reversal distance with duplicates between two genomes of interest, for which an efficient heuristic algorithm was given by introducing two new optimization problems, minimum common partition and maximum cycle decomposition. Following this approach, we have implemented a high-throughput system for assigning orthologs on a genome scale, called MSOAR, and tested it on both simulated data and real genome sequence data. Our predicted orthologs between the human and mouse genomes are strongly supported by ortholog and protein function information in authorative databases, and predictions made by other key ortholog assignment methods such as Ensembl, Homologene, INPARANOID, and HGNC. The simulation results demonstrate that MSOAR in general performs better than the iterated exemplar algorithm of David Sankoff's in terms of identifying true exemplar genes.

Thomas J. Hudson
Influence of Human Genome Polymorphism on Gene Expression

Genetic variation, through its effects on gene expression and regulatory networks, plays an important role in determining biologic phenotypes and disease susceptibility. We undertook a large technical study of methods used to detect regulatory variation in allelic expression in heterozygous samples, by comparing the effect of repeated growths of LCLs from unrelated and related individuals, repeated RNA, repeated RT-PCR, and repeated detection. We analyzed, using two independent techniques: 1) quantitative sequencing of RT-PCR products and 2) Illumina Allele-Specific Expression assays. The latter involved the testing of 1,432 exonic SNPs from 767 genes. Our estimates of allele-specific expression differences in heterozygous samples (observed for approximately 20% of human genes), are similar in both approaches and the experimental variability is low relative to the level of relative allelic expression difference. In addition, in some cases, the pedigree information available is sufficient to reconstruct the mode of inheritance of the allele-specific expression.

We systematically studied sixty-four genes that were previously reported to display allelic differences in gene expression in lymphoblastoid cell lines (LCLs) used in the HapMap project. Given the hypothesis that these would be due to common cis-acting alleles, we used HapMap genotype data to search for haplotypes

at or flanking the genes that are associated with expression. We identified sixteen loci (25%) harbouring a common haplotype that affects total expression of a gene, and a further seventeen (27%) with evidence of a haplotype affecting relative allelic expression in heterozygote samples.
Altogether, we demonstrated that assessing allele-specific expression is a quick and powerful method to detect, and discriminate, several genetic and epigenetic mechanisms of gene expression regulation including imprinting, X-inactivation, random monoallelic expression and alternative splicing. We are now analyzing with this method all genes in ENCODE regions, chromosome 21 and chromosome 22, as a prelude to whole genome studies of cis-acting regulatory variation.

Liqing Zhang
A Roadmap of Tandemly Arrayed Genes in the Mammalian Genomes

Tandemly arrayed genes (TAG) play an important functional and physiological role in the genome. Most previous studies have focused on individual TAG families in a few species, yet a broad characterization of TAGs is not available. Here we identified all the TAGs in the genomes of human, chimp, mouse, and rat and performed a comprehensive analysis of TAG distribution, TAG sizes, TAG gene orientations and intergenic distances, and TAG gene functions. TAGs account for about 14-17% of all the genomic genes and nearly one third of all the duplicated genes in the four genomes, highlighting the predominant role that tandem duplication plays in gene duplication. For all species, TAG distribution is highly heterogeneous along chromosomes and some chromosomes are enriched with TAG forests while others are enriched with TAG deserts. The majority of TAGs are of size two for all genomes, similar to the previous findings in C. elegans, Arabidopsis thaliana, and Oryza sativa, suggesting that it is a rather general phenomenon in eukaryotes. The comparison with the genome patterns shows that TAG members have a significantly higher proportion of parallel gene orientation in all species, corroborating Graham?s claim that parallel orientation is the preferred form of orientation in TAGs, and moreover, TAG members with parallel orientation tend to be closer to each other than the overall neighboring genomic genes with parallel orientation. The analyses Evol. of GO function indicate that genes with receptor or binding activities are significantly represented by TAGs. Computer simulation reveals that random gene rearrangements have little effect on the statistics of TAGs for all genomes. Finally, it is noteworthy gene family sizes are significantly correlated with the extent of tandem duplication, suggesting that tandem duplication is a preferred form of duplication, especially in large families.

Lars Feuk
Identification of structural variation in the human genome

During the last few years it has become apparent that copy number variation (CNV) and other types of structural variation are common features of the human genome. Structural variation has been shown to be functionally important but has yet to be fully ascertained. We have used two different, but complementary, approaches to characterize structural variation in the genome. Both approaches are based on comparing genomes, one being purely experimental and the other computational.

Using an experimental approach, we have constructed a first-generation map of CNVs in the human genome. The DNA samples chosen for the study were 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap collection). DNA from these samples was screened using two complementary technologies: single nucleotide polymorphism (SNP) genotyping arrays and clone based comparative genomic hybridization arrays. A total of 1,447 copy number variable regions covering 360 megabases were identified in these populations. These CNV regions contain hundreds of genes, disease loci, functional elements and segmental duplications. CNV genotype data could be extracted for many of the regions and was used to investigate the linkage disequilibrium between CNVs and SNPs. This first-generation map of CNVs in the human genome will provide a framework for future studies aiming to correlate copy number variation with human phenotypic traits.

Another approach to identify structural variation is through computational alignment of sequences obtained from different DNA sources. The most sensitive method for identifying all variation existing between two DNA donors is through direct comparison of accurately completed sequence assemblies of the genomes

under study. We developed a new algorithm and applied it for comparison of two existing assemblies, namely, Celera's R27c compilation and the human Build 35 reference sequence. Collectively, we identified megabases of sequence being present, absent, or polymorphic between the two assemblies. Using a combination of literature, database and experimental analyses, we validated 155 structural variants and more than 1.5 million SNPs. In some cases the differences were simple insertion and deletions, but when CNVs, segmental duplications, or repetitive DNA were involved the relationship was more complex. Our results highlight the need for comprehensive annotation strategies in order to fully interpret data from CNV scanning and personalized sequencing projects. The results of the studies presented here provide further support for the importance of structural variation in the human genome and highlight the value of using multiple complementary approaches to identify structural variation.