Function and Evolution of Regions Bound by Drosophila Transcription Factors

Michael Eisen, University of California, Berkeley

Identifying the genomic regions bound by sequence specific regulatory factors is central both to deciphering the complex DNA cis-regulatory code that controls transcription in metazoans and to determining the range of genes that shape animal morphogenesis. We have used whole-genome tiling arrays to map sequences bound in Drosophila melanogaster embryos by the six maternal and gap transcription factors that initiate anterior-posterior patterning. We find that these sequence-specific DNA binding proteins bind with quantitatively different specificities to an overlapping set of several thousand genomic regions in blastoderm embryos. The more highly-bound regions include all of the over forty well-characterized enhancers known to respond to these factors as well as several hundred putative new cis-regulatory modules clustered near developmental regulators and other genes with patterned expression at this stage of embryogenesis. In addition to these highly-bound regions, there are several thousand regions that are reproducibly bound at lower levels. However, these poorly-bound regions are, collectively, far more distant from genes transcribed in the blastoderm than highly-bound regions and are preferentially found in protein-coding sequences. We have extensively analyzed the evolution of recognition sequences in these bound regions and find little evidence for their preferential conservation. Together these observations suggest that many of these poorly-bound regions are not involved in early-embryonic transcriptional regulation and may be non-functional. I will propose that the pervasive view amongst both experimental and computational biologists that most protein-DNA interactions observed in vivo are functional is wrong.

Quantifying Gene Gain and Loss in Mammals

Matthew Hahn, Indiana University

Newly duplicated genes have one of two eventual fates: loss or maintenance by natural selection. Though loss is considered to be the much more likely outcome, the evolutionary modes by which duplicates are maintained have attracted much more theoretical and empirical attention. In this talk I address both of these outcomes. Using the genomes of multiple mammalian species, we are able to quantify the amount of gene gain and loss and ask whether natural selection has played a role in these changes. We have developed a novel method that allows estimation of rate heterogeneity in gain and loss among lineages; using this method we find that the rate of gene turnover in primates is more than 2.5X faster than in other mammals and may be due to both mutational and selective forces. By reconciling the gene trees for all of the gene families included in the analysis, we are able to independently verify the numbers of inferred duplications. We also use two methods based on the genome assembly of rhesus macaque to further verify our results. Our analyses identify several gene families that have expanded or contracted more rapidly than is expected even after accounting for an overall rate acceleration in primates, including brain-related families that have more than doubled in size in humans.

Looking only at duplicates specific to the human lineage, we identify genes that have undergone adaptive natural selection since our split with chimpanzees. We also examine the frequency of gene loss in these lineages, both among duplicate genes and single-copy genes. Gene loss occurs at high rates in all species, though we are able to demonstrate selection against the loss of single-copy genes.

Multiple Genetic Differences in Sialic Acid Biology Between Humans and Great Apes

Ajit Varki, University of California, San Diego

As “nothing in biology makes sense, except in the light of evolution” (Dobzhansky 1973), it follows that understanding human evolution will shed much light on the origins, mechanisms and therapy of human disease. One powerful approach to human evolution is the genetic one. When comparing individual protein sequences, humans are remarkably similar to the “great apes” (chimpanzees, bonobos, gorillas and orangutans), our closest evolutionary relatives. Indeed, studies of these relationships have resulted in a reclassification of the entire clade as “hominins”, a term previously reserved for humans and their fossil ancestors. Despite these similarities, there are remarkable differences between humans and “great apes” in the incidence and severity of many major diseases, and some of these differences should have genetic explanations. We have discovered multiple genetic and biochemical differences between humans and great apes, in relation to cell surface sugars called Sialic acids (Sias) and in a family of receptors called Siglecs (Sia-recognizing Ig superfamily lectins). One major kind of Sia called Neu5Gc is widely expressed in mammals including great apes, but not in humans. This dramatic change in the human “sialome” resulted from an inactivating mutation in the human CAMH gene, mediated by an Alu replacement ~3 million years ago. This apparently set in motion a series of additional human-specific changes in the biology of single-copy genes. Our analyses have so far been found to affect >10 genes. In particular, we find multiple human-specific differences in a family of molecules called the Siglecs (Sialic Acid-recognizing Immunoglobulin Family Lectins). Since <60 genes are involved overall in synthesis, recognition and turnover of Sias, these changes are likely related to one another in evolutionary history and potentially significant for human evolution, biology and disease. Indeed, there are implications for unique features of the human phenotype, ranging from susceptibility or resistance to some microbial pathogens; effects on the reactivity of the immune system; unusual aspects of human parturition, and the expression of Neu5Gc as a “foreign” antigen in cancers. Neu5Gc can also be metabolically incorporated from animal-derived culture materials and feeder layers into biotherapeutic molecules and cellular preparations - and into intact humans from dietary sources, particularly red meat and milk products. Meanwhile, all humans have significant levels of circulating antibodies directed against Neu5Gc-containing epitopes. This has implications for human reactions to biotechnology products, and also for com-
mon human diseases associated with chronic inflammation, such as atherosclerosis and cancer.

**Accelerated and Biased Nucleotide Evolution in the Human Genome**

Katherine S. Pollard, University of California, Davis

Comparative genomics allows us to search the whole human genome for examples of lineage-specific evolution. Using a novel likelihood ratio testing approach, we recently identified 202 Human Accelerated Regions (HARs) that were extensively changed in the last ~6 million years since divergence from our common ancestor with chimpanzee, but are highly conserved in other species and thus are likely to be functional. The HARs are mostly non-coding sequences, and the set of genes near HARs is enriched for transcription factors, suggesting a role for HARs in the evolution of human gene regulation. We will describe a few of the most intriguing HARs before turning to a curious observation about the HAR sequences: the most accelerated regions show a striking bias for AT to GC (“weak-to-strong”) nucleotide substitutions. To investigate whether this association between rate of substitutions and nucleotide bias is a genome-wide phenomenon, we quantified substitution density and bias across the human and chimp genomes. While there is no weak-to-strong bias overall, clusters of nearby substitutions (5 or more within 300bp) are highly biased. Interestingly, human polymorphisms do not show the same pattern, suggesting a fixation (rather than mutation) bias. We found a strong correlation between nucleotide bias and male recombination rates. This observation will be used to speculate about the cause of rapid, biased evolution in the primate genome and to date the chromosome fusion that formed human chr2.

**Genomics and Bioinformatics: Applications to Conservation Efforts for Endangered Species**

Oliver A. Ryder, Conservation and Research for Endangered Species, Zoological Society of San Diego

Whole genome sequencing (WGS) efforts for vertebrate species are rapidly expanding. Funding for these efforts is largely derived from the interest in identifying the functional elements and annotation of the human genome. Human and mouse are genomes are considered completely sequenced; an additional 32 mammalian species are in assembly and 34 mammalian WGS projects are in the sequencing phase according to NCBI in August, 2007. Many organisms being subjected to WGS are non-model organisms for which a relative paucity of data about their physiology, genetics, and reproduction is available. Yet, comparative studies of genome sequences are leading to a new understanding of the biology of mammals and other vertebrates, and are providing ongoing insights into the fundamental aspects of the organization and evolution of biological systems.

While the technology for DNA sequencing and harvesting information vital to understanding the biology of species is rapidly accumulating, biodiversity losses are also increasing at unprecedented rates. Genetic resource collections now provide essential tools that facilitate scientific studies. Complete WGS sequenc-
transcripts, we have compared their sequences and structures to rebuild the molecular processes responsible for their progressive diversification. We drew particular attention to the duplication of small internal segments that occurred asymmetrically between paralogs. Such events can indeed have great impact on both gene structure and protein sequence. We have collected cases of internal duplications of coding regions that lead to the acquisition of novel and functionally important amino acid repeats. We have also identified genes whose internal duplication introduces frameshifts in the encoded sequence resulting in completely novel proteins. We have observed that internal duplications can modify gene splicing by introducing novel splicing sites that affect exon-intron boundaries. All those cases will be reviewed and the likely implications of the described modifications on gene functionality will be discussed.

This work was supported by the NUSUG grant of the Italian Association for Cancer Research (AIRC).

JOINT CG-CCB SESSION

Keynote

Dynamic Duplications in the Human Genome

Barbara J. Trask, Fred Hutchinson Cancer Research Center

At least 5% of the human genome arose through segmental duplication during primate divergence. We have focused on the many very recent interchromosomal duplications and transfers concentrated at the very ends of human chromosomes. Most human subtelomeres are patchworks of sequence blocks shared by different sets of other chromosome ends; the blocks range in size from ~3 kb to 50 kb. Each sequence block varies in copy number and chromosomal distribution among humans, and thus most subtelomeres show a high degree of variation between and within species. At least half of the sequence in these regions appears to have been generated through such exchanges since human and chimpanzee diverged. Since some changes are very recent, different chromosomal ends can be more similar to each other than are alleles of the same chromosome. Detailed computational analyses of the breakpoints of homology between different chromosome sets reveals that this patchwork structure arose through a multiplicity of chromosomal translocations mediated primarily by non-homologous end joining of double-strand breaks. These regions are also subject to frequent homology-based exchange events, as evidenced by incongruent phylogenetic relationships of neighboring sections. Subtelomeres also incur 100-1000x more sister chromatid exchanges per bp than do other regions of the genome; SCEs are a measure of DNA damage and repair. Despite, or perhaps as a consequence of, the evolutionary dynamics of subtelomeres, these regions contain genes, ranging from odorant receptor genes to highly conserved essential genes. Thus, subtelomeric dynamics and variation might have phenotypic consequences. Clusters of odorant receptor genes elsewhere in the genome have also experienced a high degree of duplication, deletion, and other structural changes during mammalian evolution, leading to highly divergent chemosensory receptor repertoires in different mammals. The high sequence identity of some of OR gene clusters makes them susceptible to deleterious rearrangement with pathological consequences. OR regions are also highly enriched in those portions of genome subject to copy number variation in humans. One subtelomeric OR present in 6 to 11 apparently functional copies in different humans is an extreme example of this type of normal variation. However, enrichment of ORs in regions of copy-number variation is probably evolutionarily neutral (not selected for or against) and merely a consequence of selection against dosage changes in most other regions of the genome. This work was supported by RO1 GM57070, RO1 DC04209, and T32 HG00035.

KEYNOTE ABSTRACTS

Computational Cancer Biology

From Genomic Profiling to Biological Mechanism in Cancer Research: The Road to Therapeutic Opportunity

Paul Meltzer, National Cancer Institute

Abstract TBA

The Cancer Genome: Early Lessons

Richard Wilson, Washington University at St. Louis

During the last few years, using the newly minted reference sequence of the human genome as a foundation, initial large-scale efforts have begun to unravel the complex mutational spectrum that underlies the various forms of cancer. Using high-throughput DNA sequencing, array-based methods, and several emerging technologies, these efforts already have produced interesting and medically applicable results. In this lecture, I will review some of the early findings, present our current work, and discuss the prospects of future genome-based cancer research.

Systems Biology of Breast Cancer: Molecular Profiling at the DNA, mRNA and miRNA Level — Relevance for Prognostication and Therapy Prediction

Anne-Lise Borresen-Dale, Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Centre (Oslo, Norway)

Microarray technologies, applied to the study of DNA/RNA, can be used to portray a tumor’s detailed phenotype in its unique context, and to generate molecular signatures that will improve our understanding of the causes and progression of the disease, for the discovery of new molecular markers, for therapeutic intervention and for developing new prevention strategies. We have performed expression studies and genome wide copy number analyses of more than 500 breast carcinomas of different stages and histological types aiming at novel tumour classification that can predict survival and treatment response. The expression patterns observed provided a remarkably distinctive molecular portrait of each tumour. The tumours can be classified into 5 novel subtypes that were distinguished by pervasive differences in their gene expression patterns, and was associated with different
outcomes for the patients. In a study where DTC in bone marrow were evaluated at the time of diagnosis, presence of DTC identified a subgroup of luminal A patients with particular poor outcome. This was not apparent for other tumor subtypes. Gene expression profiles associated with DTC and with systemic relapse for luminal A patients were identified. This suggests that DTC in BM differentially distinguishes clinical outcome in patients with luminal A type tumors and that DTC-associated gene expression analysis may identify genes of potential importance in tumor dissemination.

Microarray-based Comparative Genomic Hybridization (array-CGH) provides means to quantitatively measure DNA copy aberrations (CNAs). Such aberrations occur frequently in breast cancer and define key pathogenetic events. In a genome-wide array CGH survey of CNAs using we could link distinct cytoband loci harboring CNA to specific clinicopathological parameters. Notably, distinct spectra of CNAs were found to underlie the different subtypes of breast cancer recently defined by expression-profiling, implying these subtypes develop along distinct genetic pathways. Higher numbers of gains/losses were associated with the “basal-like” tumor subtype, while DNA amplification was more frequent in “luminal-B” subtype tumors, suggesting also that distinct mechanisms of genomic instability might underlie their pathogenesis.

We have also investigated DNA copy number profiles versus RNA expression patterns in 20 human breast tumors using WG Agilent arrays. Approximately 10% of the total variation in expression could be explained by CNAs. Highly significant correlations were observed within a set of 420 genes/geneloci. In particular high correlations were seen within clusters of gene/geneloci on chromosome region 6q, 8q, 16q and 17q. Trans-regulation between CNAs and expression was explored selecting genes/loci highly correlated on the DNA and RNA level. Several different genomic regions were identified with changes in the DNA profiles correlating to significant variations in gene expression profiles in other genomic loci. These CNA regions are likely to harbor genes that control the regulation of a number of other genes. Our findings of direct or in-trans correlations of gene expression versus copy numbers offers the opportunity to identify novel interaction and to investigate their biological role and possible therapeutic application.

Recently a new class of regulatory genes, termed microRNAs (miRNA) has been identified in humans. These are non-coding, meaning that they do not encode for proteins but produces small single stranded RNAs, 20-22 nt long that interfere with mRNA either by preventing translation or promoting degradation. They have been linked to numerous biological processes, including cell proliferation and cell death during development, stress resistance, and fat metabolism. Microarrays with the approximately 600 known miRNA to date have been made by Agilent, and we have in a pilot project analyzed 20 breast tumors using these arrays. The expression profile of approximately 50 miRNA was able to classify the basal tumors vs the luminal with close to 100% predictability. This has set the stage for including miRNA expression analyses in most of our tumor sets analyzed by expression. We envision that the combined miRNA-mRNA expression profile will, in addition to contribute to our understanding of the development of the different subclasses, be good predictors of therapy response and guide us in identifying new targets for therapy.

In order to elucidate, in a larger scale, the effect of the genetic component on the expression in tumor tissue, candidates SNPs were analyzed for associations to an unselected whole genome pool of tumor mRNA transcripts in 50 unrelated patients with breast cancer. SNPs were selected from 203 candidate genes of the reactive oxygen species (ROS) pathway. Every gene has a number of SNPs, from a few to a hundreds, dependent on the size and degree of conservation. A general statistical framework was developed for the simultaneous analysis of gene expression data and SNP genotype data measured for the same cohort. It revealed significant associations between subsets of SNPs and transcripts, shedding light on the underlying biology. SNPs in important regulators such as EGF, IL1A, MAPK8, XPC, SOD2 and ALOX12 were associated with the expression patterns of a significant number of transcripts, indicating the presence of regulatory SNPs in these genes. SNPs were found to act in-trans in a total of 115 genes. Of these, SNPs in 43 of these 115 genes were found to act both in-cis and in-trans. Finally, subsets of SNPs that share significantly many common associations with a set of transcripts (biclusters) were identified. The subsets of transcripts that are significantly associated to the same set of SNPs or to a single SNP were shown to be functionally coherent in GO and pathway analyses and co-expressed in other independent data sets, suggesting that many of the observed associations are within the same functional pathways.

All the above findings set the stage for future studies aimed at identifying early markers for breast cancer, and genotypes influencing the specific patterns of gene activations that predict important clinical features, like sensitivity to specific therapies and metastatic potential. However, despite the optimistic prospective for molecular profiling to change cancer management, there are several challenges that need to be addressed: Statistical and analytical handling of microarray data is a critical issue which needs further attention. Combining analytical strategies with functional hypotheses are needed to reach a deeper understanding of the underlying biological processes and mechanisms controlling tumor behavior. Finally, large prospective trials are needed to validate the numerous gene signatures that have been identified in various studies, and in time we hope to arrive at molecular signatures that can be used in clinical practice.

**Causes and Consequences of Genomic Instability in Cancer**

**Donna Albertson**, University of California, San Francisco

Genomes of solid tumors are characterized by gains and losses of regions, which may contribute to tumorigenesis by altering gene expression. Application of genome-wide array comparative genomic hybridization (CGH) in breast and oral cancer has revealed tissue specific genome alterations, variation in the numbers and types of aberrations, and discrimination of tumor subtypes based on the spectrum of copy number alterations. The underlying causes of genome instability have been investigated in tumors and model systems with respect to contributions from both deregulated expression of genes that maintain genome sta-
Comparative Genomics

Max Alekseyev, University of California, San Francisco

Multi-break rearrangements break a genome into multiple fragments and further glue them together in a new order. While 2-break rearrangements represent standard reversals, fusions, fissions, and translocations operations; 3-break rearrangements are a natural generalization of transpositions and inverted transpositions. Multi-break rearrangements in circular genomes were studied in depth by Alekseyev and Pevzner, 2007 and were further applied to the analysis of chromosomal evolution in mammalian genomes in another paper by Alekseyev and Pevzner, 2007. In this paper, we extend these results to the more difficult case of linear genomes. In particular, we give lower bounds for the rearrangement distance between linear genomes and use these results to analyze comparative genomic architecture of mammalian genomes.

Learning Gene Regulatory Networks via Globally Regularized Risk Minimization

Yuhong Guo, University of Alberta; and Dale Schuurmans

Learning the structure of a gene regulatory network from time-series gene expression data is a significant challenge. Most approaches proposed in the literature to date attempt to predict the regulators of each target gene individually, but fail to share regulatory information between related genes. In this paper, we propose a new globally regularized risk minimization approach to address this problem. Our approach first clusters genes according to their time-series expression profiles – identifying related groups of genes. Given a clustering, we then develop a simple technique that exploits the assumption that genes with similar expression patterns are likely to be co-regulated by encouraging the genes in the same group to share common regulators. Our experiments on both synthetic and real gene expression data suggest that our new approach is more effective at identifying important transcription factor-based regulatory mechanisms than the standard independent approach and a prototype based approach.

Evolution of Tandemly Arrayed Genes in Multiple Species

Mathieu Lajoie, University of Montreal; Denis Bertrand and Nadia El-Mabrouk

Tandemly arrayed genes (TAG) constitute a large fraction of most genomes and play important biological roles. They evolve through unequal recombination, which places duplicated genes next to the original ones (tandem duplications). Many algorithms have been proposed to infer a tandem duplication history for a TAGs cluster in a single species. However, the presence of different transcriptional orientations in most TAGs clusters highlights the fact that processes such as inversions also contribute to their evolution. This makes those algorithms unsuitable in many cases. To circumvent this limitation, we proposed in a previous work an extended evolutionary model which includes inversions and presented a branch-and-bound algorithm allowing to infer a most parsimonious scenario of evolution for a given TAGs cluster.

Here, we generalize this model to multiple species and present a general framework to infer ancestral gene orders that minimize the number of inversions in the whole evolutionary history. An application on a pair of human-rat TAGs clusters is presented.

A Pseudo-Boolean Programming Approach for Computing the Breakpoint Distance between Two Genomes with Duplicate Genes

Sébastien Angibaud, Guillaume Fertin, Irena Rusu, Annelyse Thévenin, Université Paris-Sud; and Stéphane Vialette

Comparing genomes of different species has become a crucial problem in comparative genomics. Recent research has resulted in different genomic distance definitions: number of breakpoints, number of common intervals, number of conserved intervals, Maximum Adjacency Disruption number (MAD), etc. Classical methods (usually based on permutations of gene order) for com-
puting genomic distances between whole genomes are however seriously compromised for genomes where several copies of the same gene may be scattered across the genome. Most approaches to overcoming this difficulty are based on the exemplar method (keep exactly one copy in each genome of each duplicated gene) and the maximum matching method (keep as many copies as possible in each genome of each duplicated gene). Unfortunately, it turns out that, in presence of duplications, most problems are NP-hard, and hence several heuristics have been recently proposed. Extending research we initiated in JCB, 2007, we propose in this paper a novel generic pseudo-boolean approach for computing the exact breakpoint distance between two genomes in presence of duplications for both the exemplar and maximum matching methods. We illustrate the application of this methodology on a well-known public benchmark dataset of γ-Proteobacteria.

Reconstructing an Inversion History in the Anopheles gambiae Complex

Ai Xia, Maria V. Sharakhova and Igor Sharakhov, Virginia Tech

The phylogenetic relationships among the members species complexes can be inferred from the distribution of fixed inversions if outgroup arrangements are known. The Anopheles gambiae complex consists of seven African mosquito species that can be differentiated based on ten fixed inversions. However, the phylogenetic relationships among the members remain unclear. This paper demonstrates that physical maps of the outgroup species A. funestus and A. stephensi can be used for determining ancestral chromosome arrangements in the A. gambiae complex. Gene order comparisons have been performed using the Multiple Genome Rearrangements (MGR) and Sorting Permutations by Reversals and block-INterchanGes (SPRING) programs. The analysis has identified the chromosomal arrangements which are likely to be the ancestral in the complex.

Improving Inversion Median Computation using Commuting Reversals and Cycle Information

William Arndt, University of South Carolina; and Jijun Tang,

In the past decade, genome rearrangements have attracted increasing attention from both biologists and computer scientists as a new type of data for phylogenetic analysis. Methods for reconstructing phylogeny from genome rearrangements include distance-based methods, MCMC methods and direct optimization methods. The latter, pioneered by Sankoff and extended with the software suite GRAPPA and MGR, is the most accurate approach, but is very limited due to the difficulty of its scoring procedure—it must solve multiple instances of median problem to compute the score of a given tree. The median problem is known to be NP-hard and all existing solvers are extremely slow when the genomes are distant. In this paper, we present a new inversion median heuristic for unichromosomal genomes. The new method works by applying sets of reversals in a batch where all such reversals both commute and do not break the cycle of any other. Our testing using simulated datasets shows that this method is much faster than the leading solver for difficult datasets with only a slight accuracy penalty, yet retains better accuracy than other heuristics with comparable speed. This new method will dramatically increase the speed of current direct optimization methods and lets us extend the range of their applicability to organellar and small nuclear genomes with more than 50 inversions along each edge. As a further improvement, this new method can very quickly produce reasonable solutions to problems with hundreds of genes.

Recovering True Rearrangement Events on Phylogenetic Trees

Hao Zhao, Genome Institute of Singapore; and Guillaume Bourque

Given the gene-order of a set of contemporary genomes, the problem of recovering the rearrangement scenario that best explains these arrangements can be challenging even if the phylogeny of these species is known. Most of the existing methods can identify an optimal or near-optimal scenario in terms of parsimony but they cannot distinguish between reliable and putative events on the reconstructed tree. In this paper, we propose an efficient method to infer partial rearrangement scenarios consisting of only reliable ancestral events. Using simulations, we show that the approach allows the recovery of actual events with high sensitivity and specificity under both random and fragile rearrangement models. Finally, we also apply the approach to two real data sets.

Baculovirus Phylogeny Based on Genome Rearrangements

Daniel Goodman, University of California, San Diego

Deriving phylogeny of rapidly evolving viral genomes is one of the most challenging problems in evolutionary studies since gene-based phylogenies often produce conflicting evolutionary trees. Phylogenetic reconstruction methods that consider whole genomes are becoming more reliable with the increasing availability of complete genome sequences and the development of algorithms to compare entire genomes. Here we employ a gene rearrangement-based approach to study Baculovirus phylogeny. Since genome rearrangement algorithms require the set of genes shared between all genomes, the most challenging problem in analyzing rapidly evolving genomes is generating this set of genes. Indeed, there are fewer and fewer genes shared between N species as N increases. We address this challenge by iteratively considering smaller sets of related genomes to find conserved genes. Baculovirus was chosen as a test case because a large number of its constituent genomes have been sequenced and its evolutionary relationships are well studied. The resulting phylogenies show clear separation of Baculoviridae into Nucleopolyhedrovirus (NPV) and Granulovirus (GV) as well as the separation of NPVs into groups I and II. Further species separation results in phylogenetic relationships that are largely consistent with conventional gene-based approaches, with some differences that provide insight into the rearrangements of Baculoviridae genomes. Our open source software, MULGOR (MULtiple
Genome Order), which analyzes genes shared between multiple small rapidly evolving genomes, is available at http://realm.sdsc.edu/MULGOR/.

Selecting Genomes for Reconstruction of Ancestral Genomes

Guoliang Li, Jian Ma, UC Santa Cruz; and Louxin Zhang

It is often impossible to sequence all descendent genomes to reconstruct an ancestral genome. In addition, more genomes do not necessarily give a higher accuracy for the reconstruction of ancestral character states. These facts lead to studying the genome selection for reconstruction problem. In this work, two greedy algorithms for this problem are proposed and tested on computer simulation data as well as a biological example.

How to Achieve an Equivalent Simple Permutation in Linear Time

Simon Gog, University of Ulm; and Martin Bader

The problem of Sorting signed permutations by reversals is a well studied problem in computational biology. The first polynomial time algorithm was presented by Hannenhalli and Pevzner in 1995. The algorithm was improved several times, and nowadays the most efficient algorithm has a subquadratic running time (Tannier and Sagot, 2004; Tannier, Bergeron, and Sagot, 2007). Simple permutations played an important role in the development of these algorithms. Although the latest version of Tannier et al., 2007 does not require simple permutations the preliminary version of their algorithm (Tannier and Sagot, 2004) as well as the first polynomial time algorithm of Hannenhalli and Pevzner use the structure of simple permutations. However, the latter algorithms require a precomputation that transforms a permutation into an equivalent simple permutation. To the best of our knowledge, all published algorithms for this transformation have at least a quadratic running time. For further investigations on genome rearrangement problems, the existence of a fast algorithm for the transformation could be crucial. In this paper, we present a linear time algorithm for the transformation.

A Rigorous Analysis of the Pattern of Intron Conservation Supports the Coelomata Clade of Animals

Jie Zheng, National Center for Biotechnology Information, National Institutes of Health; Igor Rogozin, Eugene Koonin and Teresa Przytycka

Many intron positions are conserved in varying subsets of eukaryotic genomes and, consequently, comprise a potentially informative class of phylogenetic characters. Roy and Gilbert developed a method of phylogenetic reconstruction using the patterns of intron presence-absence in eukaryotic genes and, applying this method to the analysis of animal phylogeny, obtained support for an Ecdysozoa clade. The critical assumption in the method was the independence of the rates of intron loss in different branches of the phylogenetic. Here, this assumption is refuted by showing that the branch-specific intron loss rates are strongly correlated. We show that different tree topologies are obtained, in each case with a significant statistical support, when different subsets of intron positions are analyzed. The analysis of the conserved intron positions supports the Coelomata topology, i.e., a clade comprised of arthropods and chordates, whereas the analysis of more variable intron positions favors the Ecdysozoa topology, i.e., a clade of arthropods and nematodes. We show, however, that the support for Ecdysozoa is fully explained by parallel loss of introns in nematodes and arthropods, a factor that does not contribute to the analysis of the conserved introns. The developed procedure for the identification and analysis of conserved introns and other characters with minimal or no homoplasly is expected to be useful for resolving many hard phylogenetic problems.

A Heuristic Algorithm for Reconstructing Ancestral Gene Orders with Duplications

Jian Ma, UC Santa Cruz; Aakrosh Ratan, Louxin Zhang, Webb Miller and David Haussler

Accurately reconstructing the large-scale gene order in an ancestral genome is a critical step to better understand genome evolution. In this paper, we propose a heuristic algorithm for reconstructing ancestral genomic orders with duplications. The method starts from the order of genes in modern genomes and predicts predecessor and successor relationships in the ancestor. Then a greedy algorithm is used to reconstruct the ancestral orders by connecting genes into contiguous regions based on predicted adjacencies. Computer simulation was used to validate the algorithm. We also applied the method to reconstruct the ancestral genomes of ciliate Paramecium tetraurelia.

Parts of the Problem of Polyploids in Rearrangement Phylogeny

Chunfeng Zheng, Qian Zhu and David Sankoff, University of Ottawa

Genome doubling simultaneously doubles all genetic markers. Genome rearrangement phylogenetics requires that all genomes analyzed have the same set of orthologs, so that it is not possible to include doubled and unduplicated genomes in the same phylogeny. A framework for solving this difficulty requires separating out various possible local configurations of doubled and unduplicated genomes in a given phylogeny, each of which requires a different strategy for integrating genomic distance, halving and rearrangement median algorithms. In this paper we focus on the two cases where doubling precedes a speciation event and where it occurs independently in both lineages initiated by a speciation event. We apply these to a new data set containing markers that are ancient duplicates in two yeast genomes.
On the frequency of genome rearrangement events in cancer karyotypes

Michal Ozery-Flato and Ron Shamir, Tel Aviv University

The results of this instability can be observed in the karyotypes of many cancerous genomes, which often contain a variety of aberrations. In this study we introduce a new approach for analyzing rearrangement events in carcinogenesis. This approach builds on a new effective heuristic for computing a short sequence of rearrangement events that may have led to a given karyotype. We applied this heuristic on over 40,000 karyotypes reported in the scientific literature. Our analysis implies that these karyotypes have evolved predominantly via four principal event types: chromosomes gains and losses, reciprocal translocations, and terminal deletions. We used the frequencies of the reconstructed rearrangement events to measure similarity between karyotypes. Using clustering techniques, we demonstrate that in many cases, rearrangement event frequencies are a meaningful criterion for distinguishing between karyotypes of distinct tumor classes. Further investigations of this kind can provide insight on the scenarios by which particular cancer types have evolved.

Analyzing Matched Differential Gene Expression with an Application to Cancer vs. Normal MicroRNA Data

Anya Tsalenko, Roy Navon, Israel Steinfeld, Hui Wang, Robert Ach, Bo Curry, Laurakay Bruhn, Amir Ben-Dor and Zohar Yakhini, Agilent Technologies

Methods for identifying differentially expressed genes are central in routine analysis of classified gene expression data. Variations amongst individuals is a significant confounding factor in expression profiling studies that are typically limited to a small number of individuals and where thousands of genes are typically measured. It is often the case that clinically or biologically meaningful phenomena are associated with some differential expression. The true character of the difference in expression signatures can be obscured by inter-patient variation. We present methods to overcome this in studies where samples of all or most classes under consideration were measured for each participating patient. Our methods do not make model or sample size assumptions. We discuss appropriate gene scores and how to compute them; we develop statistics for assessing the results; we conclude by demonstrating our methods in analyzing miRNA expression in matched tumor and normal samples.

A Systems Biology Approach to the Identification of Causal Somatic Mechanisms in Cancer Phenotypes

Kartik Mani, Celine Lefebvre, Kai Wang, Wei Keat Lim, katia Basso, Riccardo Dalla-Favera and Andrea Califano, Columbia University.

Cancer is a heterogeneous somatic disease, originating in different tissues and involving a number of distinct molecular pathways. While its complexity presents an enormous challenge, the availability of high-throughput biological data is driving a paradigm shift in oncology research. Integrative computational approaches that leverage the wealth of this molecular information are increasingly being applied to decipher oncogenic mechanisms. Many methods have been successful at identifying diagnostic and prognostic cancer signatures from data such as genome-wide expression profiles. However, they fail to provide a full mechanistic understanding of the pathologic transition, as causal chromosomal aberrations and the specific downstream affected pathways cannot be clearly inferred. Several methods have addressed the problem of identifying causal genetic events in cancer phenotypes by integrating genetic and genomic profiles. However, these also suffer from the limitation of a being applied in a gene-centric context. In this study, we argue that new insight into tumorigenesis can be derived from a systems biology perspective. We first generate a comprehensive network of regulatory, physical and signaling interactions for human B cells using a Bayesian evidence integration framework. We then identify interactions that show aberrant behavior in specific phenotypes against the average using a large compendium of microarray expression profiles. Finally, we score genes based on their proximity to these “dysregulated” interactions. We show that for this well annotated phenotypes (3 cancer, 1 normal), this method correctly identifies the causal gene reported in the literature in the top 0.3% of all candidates, as well as key downstream effectors.

Primer Selection Methods for Detection of Genomic Inversions and Deletions via PAMP

Bhaskar DasGupta, Jin Jun and Ion Mandoiu, University of Connecticut

Primer Approximation Multiplex PCR (PAMP) is a recently introduced experimental technique for detecting the location of large-scale structural variations such as inversions and deletions in cancer genomes. In this paper we give integer linear programming formulations for the problem of selecting sets of PAMP primers that minimize the probability of failure to detect inversions or deletions present in a given genomic interval.

Computational Structural Proteomics of Protein Interactions Determines Specificity Signatures of Cancer Drugs

Gennady Verkhivker, University of Kansas

The emerging insights into kinase function and evolution combined with a rapidly growing number of crystal structures of protein kinases complexes have facilitated a comprehensive structural bioinformatics analysis of sequence-structure relationships in determining the binding function of protein tyrosine kinases.

We have found that evolutionary signal derived solely from the tyrosine kinase sequence conservation can not be readily translated into the ligand binding phenotype. However, fingerprinting
ligand–protein interactions using in silico profiling of inhibitor binding against protein tyrosine kinases crystal structures can detect a functionally relevant kinase binding signal and reconcile the existing experimental data. In silico proteomics analysis unravels mechanisms by which structural plasticity of the tyrosine kinases is linked with the conformational preferences of cancer drugs Imatinib and Dasatinib in achieving effective drug binding with a distinct spectrum of the tyrosine kinase. A comprehensive study of evolutionary, structural, dynamic and energetic aspects of tyrosine kinases binding with clinically important class of inhibitors provides important insights into mechanisms of sequence–structure relationships in the kinome space and molecular basis of functional adaptability towards specific binding.

**Factors Influencing the Fate of Mammalian Gene Duplicates**

Paul Ryvkin, Jin Jun, Edward Hemphill and Craig Nelson, University of Connecticut

What causes duplicate genes to follow different or similar evolutionary trajectories? Here we show the relative contribution of several duplication mechanisms to mammalian gene families. Additionally, we note that there is a bias towards the production of tandem gene arrays when genes duplicate multiple times. In order to examine the distinct consequences of each mechanism we examine evolutionary rate asymmetry of gene pairs. We show that relocated and retrotransposed duplications yield genes whose sequences evolve more asymmetrically than tandem duplicates. Finally we show that among distant segmental duplicates, disruptions in flanking regions correlate with an increase in substitution rates and a relaxation of selective constraint, providing evidence that abrupt changes in cis-regulatory regions as well as variations in local mutation rates can have profound effects on protein evolution in duplicate gene copies.

**Spatial Fluctuation of Recombination Rates in the HIV Genome: A Computational Model Identifies Hotspots**

Misha Rajaram, Vladimir Minin, Marc A. Suchard and Karin Dorman, Iowa State University

Coinfection of a single host cell with two parental HIV strains may produce recombinant viruses upon template switching by the replication machinery. A multiple change point, Bayesian model was used to infer inter-subtype recombination occurring during the worldwide epidemic of HIV. An initial screening produced 528 unique recombinants representing 985 unique crossover points. Recombination rate was not uniform across the HIV-1 genome. A modest recombination hotspot appeared in the p24 region of the gag gene, corroborating a previous finding. Recombination was considerably more frequent throughout the env reading frame. However, the dominant hotspot for recombination occurred in the Reverse transcriptase gene, coincident with a cluster of drug resistance mutations and raising the possibility of in vivo selection for recombinant genomes.

**Selection on Genome Arrangement in a Recently Emerged Bacterial Pathogen**

Aaron Darling, University of Queensland; István Miklós and Mark A. Ragan

We apply Bayesian methods to infer inversion phylogeny and selection on genome arrangement in eight *Yersinia* genomes. A Mauve genome alignment of the eight *Yersinia* reveals 78 Locally Collinear Blocks (LCBs), which we project to a signed gene-order permutation matrix. We then apply modified BADGER software which uses MCMCMC (Metropolis-Coupled Markov-chain Monte Carlo) to sample the joint posterior distribution of phylogenetic trees, inversion rates, and inversion histories. BADGER assumes *a priori* that all inversions are equally likely, and parsimonious histories are more likely.

The posterior sampling contains seven parsimonious tree topologies with 79 inversions each. Sampled topologies agree with results suggested by SNP patterns. Ancestral genome arrangement demonstrates strong selection for replichore balance in *Yersinia* (replication terminus positioned opposite the replication origin on the circular chromosome). Breakpoint reuse analysis reveals hotspots of genome rearrangement near the origin, with a “non-divisible zone” near the terminus. We also discover that inversions acting within a single replichore are much shorter than other inversions.

**Deciphering the Contribution of Cis-Regulatory Variants to Gene Expression Patterns**

Joshua Rest, University of Chicago; Kevin Bullaughey, Geoffrey Morris and Wen-Hsiung Li

Gene expression is controlled by transcription factors that bind to a variable family of motifs in a gene’s promoter; however the role of variability within motif families has not been systematically studied. Deciphering the function of these variable positions is important for understanding the complex cis-regulatory code that underlies the physiology, ecology and evolution of gene expression. Using a computational approach, we show that functional variants of these motifs among genes within the yeast genome are associated with condition-specific differential expression. The regulatory consequences of these variants are often conserved across different yeast species, and their evolution is associated with change in gene expression. The power of binding site variants to predict gene expression is tested by experimentally switching between binding site variants through site-directed mutagenesis in yeast.

**Coalescent Consequences for Consensus Cladograms**

James Degnan, University of Michigan; Michael Degiorgio, David Bryant and Noah Rosenberg.

To investigate the theoretical properties of consensus trees obtained from gene trees evolving at different loci, we construct
consensus trees from independent gene trees that occur in proportion to their probabilities from coalescent theory. We consider majority-rule, rooted triple (R*), and greedy consensus asymptotically as the number of loci approaches infinity and for finite samples. We investigate the effects of branch lengths on the species tree and find different consistency results for the three methods: majority-rule consensus trees can be increasingly likely to be unresolved, although not misleading; greedy consensus trees can be misleading for all species tree topologies with at least five taxa; and R* consensus is statistically consistent, i.e., guaranteed to return the species tree topology given enough loci, for any binary species tree topology with any set of branch lengths. We also investigate the performance of consensus trees as species tree estimators for finite numbers of loci.

Reconciliation With Non-Binary Species Trees

Maureen Stolzer, Benjamin Vernot, Aiton Goldman and Dannie Durand, Carnegie Mellon University

The previously cloned and characterized cpcA gene of Aspergillus niger is found to display the same functions of GCN4 as in Saccharomyces cerevisiae. In S. cerevisiae Gcn4p activates transcription of various amino acid biosynthetic genes, pathway specific activators, aminoacyl t-RNA and purine synthethase gene. This regulatory response is named as General Amino Acid Control (GAAC) in S. cerevisiae and cross pathway control (CPC) in filamentous fungi. In this study it is aimed to determine the mode of action of the regulation of A. niger cpcA and the transcriptional activation effects of the same gene product in response to different stress conditions. The stress conditions possible to interact with the pathways in concern are, phenyl lactic acid, rapamycin and 3-aminotriazole. The cpcA knockout strain constructed will be used in a microarray. The outcome of this microarray study will further be used to target genes that play an important role in regulation of cpcA and as well the genes directly or indirectly regulated by cpcA.

Initial Steps of In-silico-subtyping of Human Tumor Tissues as Exemplified on Toponome Profiles Characterizing Individual Renal Clear Cell Carcinomas (RCC)

Michael Kreutzer, Germany; Larisa Kiseleva, Japan AIST; Gilbert Schoenfelder, Institute for Pharmacology and Toxicology, Wuerzburg, Germany; Marcus Haemmerle, Meltec Gmb; Stephan Stoerkel, University Witten Herdecke; and Hans-Juergen Thiesen, Institute of Immunology Rostock, Germany

To obtain spatial information on cell-type specific and disease-pathway related protein expressions, multidimensional microscopic high-throughput protein colocalizations studies (toponome analysis) can be employed by performing numerous cycles of fluorescence tagging, imaging and in situ bleaching on individual tissue sections once informative antibody sets become available. Hereto, informative antibody repertoires have to be selected to determine disease pathways and computational comparisons thereof.

Material and Methods: The Swedish Protein Atlas (Release 2006-10-30) content (www.proteinatlas.org) currently displaying immunostainings of 1514 antibodies relating to 1,238 760 images including microarray tissue samples from 48 different normal tissue types, 20 different types of cancer and various cell lines were analyzed by in-depth in-silico analysis in order to determine the most informative toponome profiles and antibody sets. Thus, Pearson correlation distances were calculated of immunostainings of i. normal tissue, ii. cancer tissue and iii. cell types singly and in combination to evaluate how expressions profiles diverge from each other and to determine tumour expression profiles that are more closely or more distantly correlated with each other. Finally, immunostainings of 291 antibodies characterizing 11 individual tumor specimen of renal clear cell carcinoma (RCC) were subtype in-silico classifying various degrees of RCC tumour heterogeneities. To visualize the relationships between toponome profiles of Proteinatlas immunostainings, a minimum
spanning tree—a concept from graph theory—was employed. Pearson correlation coefficients were measured between every pair of samples using R software and the correlation coefficients translated into correlation distances. The minimum spanning tree of a full graph was computed using Kruskal’s algorithm. To obtain images of computed networks the open source software GraphViz was used.

Results and Discussion: Antibody sets showing the most divergent expression patterns within 48 tissue types 20 cancer types and numerous cell lines were assessed as well as the combination of them. Minimum spanning trees thereof were visualized based on Pearson correlation distances to determine tissues and proteins thereof that determine most divergent protein expression profiles. Worth noting, expression profiles of tumour cell lines formed one cluster and did not mingle with expression profiles of tumour and normal tissues. Most importantly, the protein expression analysis of tumour tissues in conjunction with normal tissues revealed that the extracted data sets describing tumour types are highly diverse within individual cancer types revealing the presence of great inherent heterogeneities. As a consequence of the analysis of renal clear cell carcinoma (RCC) data sets, Pearson correlation distances were established on tumour expression patterns of proteins sets derived from individual patients as indicated by gender and sex, see www.proteinatlas.org.

Perspectives: The integration of multiparametric toponome analysis on one specimen encompasses a great potential to direct the selection of more efficient therapeutic approaches just by subtyping individual tumours and determining correlation distances to relate therapeutic outcomes to disease specific pathways. Our in-silico analysis demonstrates that the Proteinatlas database is a unique reservoir to describe heterogeneities within individual tumour types in order to determine the most informative antibody sets for subtyping human tumours on the way to a more personalized medicine.

The Possibility of Personalized Oncology Using End-Sequence Profiling Combined with Next Generation Sequencing Technology

Stanislav Volik, UC San Francisco; Benjamin Raphael, Brown University; Peng Yu, Chinese National Human Genome Center; Chunjie Wu, Shandong Provincial Hospital; Guigue Huang, Yasuko Kobayashi, and Pamela Paris, UC San Francisco; Joe Gray and Jan-Fang Chen, Lawrence Berkeley National Laboratory; Pieter de Jong, Children’s Hospital, Oakland; and Hesed Padilla-Nash, National Cancer Institute

Tumor genomes can be highly rearranged and not colinear with the host genome. Recurrent genome rearrangements involve genes that are increasingly targeted by anti-tumor therapeutics. Current technologies for studying tumor genomes do not determine their structure and relate it to the underlying sequence and are consequently not well equipped to identify fusion genes and the proteins they encode. Originally End Sequence Profiling (ESP) was developed as a sequence-based method for directly determining the structure of tumor genomes, and for cloning all types of rearrangements en masse. Being sequence-based, ESP is inherently integrative, thus bringing the power of genetic analy-

sis to interpretation of complex expression microarray and proteomic data. Detection of the genomic breakpoints and fusion transcripts is based on paired-end sequencing of the genomic fragments/cDNAs and alignment of these sequences to normal reference genome. This process reveals all types of structural aberrations such as copy number changes, translocations, and inversions and in current implementation identifies BAC clones spanning the genome breakpoints that can be sequenced to identify breakpoint(s) and fusion genes. We applied ESP to the breast cancer cell lines BT474, MCF7, and SKBR3, primary tumors of the brain, breast, ovary, and a metastatic prostate tumor, creating the world’s largest collection of sequence-ready tumor breakpoints. ESP provides direct evidence for packaging of amplified DNA from multiple loci, extensive rearrangements of amplicons, molecular heterogeneity, and fusion genes. In addition, ESP can be carried out on tumor transcriptomes for large-scale identification of fusion transcripts. Recent breakthroughs in massively parallel next generation sequencing technologies (NGST) eliminate BAC/fosmid cloning and cut the cost of ESP greater than 100-fold. This means that hundreds to thousands of clinically relevant tumors can be analyzed and that NGST-based ESP may play a role in personalized genomic oncology and genomic pathology.

A Multi-Dimensional Analysis of Genes Mutated in Breast and Colorectal Cancers

Jimmy Lin and Victor Velculescu, Johns Hopkins University

A recent study of a large number of genes in a panel of breast and colorectal cancers identified somatic mutations in 1,149 genes (Sjöblom et al. 2006). To identify biologic processes affected by these genes, we examined their putative roles based on sequence similarity, membership in known functional groups and pathways, and potential interactions with other proteins. These analyses identified functional groups and pathways that were enriched for mutated genes in both tumor types. Additionally, the results pointed to differences in molecular mechanisms that underlie breast and colorectal cancers, including various intracellular signaling and metabolic pathways. These studies provide a multi-dimensional framework to guide further research and help identify cellular processes critical for malignant progression and therapeutic intervention.

Prediction of Somatic Cancer Driver Mutations in Protein Kinases

Ali Torkamani, University of California, San Diego and Nicholas Schork, Scripps Genomic Medicine

Contemporary systematic resequencing of the kinome in cancer cell lines has suggested that most somatic mutations are likely to be ‘passengers’ that do not contribute to tumorigenesis or cancer development. A challenge posed by these systematic resequencing efforts is to differentiate between ‘passenger’ and ‘driver’ mutations. Recent evidence suggests that cancer mutations contributing to tumorigenesis have characteristics similar to mutations causally related to Mendelian diseases. We have developed a support vector machine-based method specifically designed to differentiate common, non-Mendelian disease-causing from
Mendelian disease causing polymorphisms in human protein kinases. Kinases are known to be major contributors to cancer development and progression. We have applied our method to the collection of somatic cancer mutations identified by Greenman et al. Our method isolates 159 specific driver mutations out of the total catalogue of mutations, demonstrating extremely good agreement with the prediction of 158 driver mutants by statistical analysis of synonymous vs. nonsynonymous mutations. Furthermore, sequence and structural analysis, as well as our analysis of the frequency at which predicted driver mutations are observed in the Cosmic database, strongly suggest the identified mutations are, in fact, true cancer drivers and not passengers.

**Predicting the Functional Impact of Somatic Mutations in Protein Kinases Using Evolutionary Constraints**

Eric Scheeff, Salk Institute; Erin Pleasance, Wellcome Trust Sanger Institute, UK; Yufeng Zhai, Patrick Schacht and Michael Dacre, Salk Institute; Ali Torkamani and Nicholas Schork, UC San Diego; Michael Stratton and Andrew Futreal, Wellcome Trust Sanger Institute; and Gerard Manning, Salk Institute

Protein kinases are essential regulators of cell growth, death, migration and differentiation, and have been intensively studied as key drivers in oncogenesis. Approximately 120 of the 518 human kinases have been implicated in cancer, and kinase inhibitors are emerging as a major new class of cancer therapies. Large-scale studies have found hundreds of non-synonymous point mutations in kinases in a variety of tumor types, but the majority of these mutations are predicted to be passenger mutations, with minimal impact on cancer progression. Therefore, it is important to single out the likely driver mutations for better diagnosis and therapeutic development. The evolutionary sequence constraints seen in orthologous proteins provide valuable signals as to the likely impact of specific human sequence mutations on subsequent protein function. Mutations that are acceptable to protein function will have been retained by the evolutionary process, but those with deleterious impact will have been removed by purifying selection. We generated high-quality predicted sequences for all human protein kinase orthologs for 5 fishes, frog, chicken, and 8 mammals. We used these aligned ortholog sets to generate a score for the likely functional impact of any mutation on each residue of almost every human protein kinase. These scores correctly predict that the majority of known disease mutations are functionally significant, and that the majority of common SNPs are benign.

Comparison of more than 500 somatic cancer mutations from a kinome-wide screen with a simulated set of random mutations shows that both sets are predicted to have a much greater functional impact than SNPs. This pattern indicates that both driver and passenger mutations are likely to have a significant impact on protein function (though only drivers are important for tumor progression). While no clear separation between driver and passenger mutations was seen, this method allows a prioritization of mutations in terms of likely functional impact and promises to be an important part of a wider predictive method for somatic mutation annotation.

**Statistical Multi-Experiment Analysis of Non-Discretized Array CGH Data**

Christiaan Klijn, Delft University of Technology; Henne Holstege, Netherlands Cancer Institute; Jeroen de Ridders, Delft University of Technology; Xiaoling Liu, Netherlands Cancer Institute; Marcel J.T. Reinders, Delf University of Technology; Jos Jonkers, Netherlands Cancer Institute; and Lodewyk Wessels, Netherlands Cancer Institute

Tumor formation is in part driven by copy number alterations (CNAs), which can be measured using array Comparative Genomic Hybridization (aCGH). Multi-experiment analysis of array comparative genome hybridization data from tumors allows discovery of recurrent copy number alterations (CNAs) that are potentially causal to cancer development. Until now, multi-experiment aCGH data analysis has been dependent on discretization of measurement data to a gain (1), loss (-1) or no-change (0) state. Valuable biological information is lost when a complex, heterogeneous system such as a solid tumor is reduced to these three states. We have developed Kernel Convolution - a Statistical Method for Aberrant Region DeTection (KC-SMART), a new approach which inputs non-discretized log2 aCGH data to identify regions that are significantly aberrant across an entire tumor set.

We use kernel convolution to generate a Kernel Smoothed Estimate (KSE) of CNAs across the genome, aggregated over all tumors. This allows us to account for the strength of a probe’s log2 signal and the distribution of the signal across multiple tumors. It also allows us to incorporate the local genomic environment of a probe by smoothing across a neighborhood of multiple probes and compensating for non-uniform probe spacing on the aCGH platform. By varying the width of the kernel function, a scale space is created which enables identification of aberrations of all sizes. Statistical significance is determined by permutation and results are corrected for multiple testing to control the false discovery rate.

In an analysis of 89 human sporadic breast tumors our method identified the known loss of a region of chromosome 5q with triple negative receptor status (ER, PR and ERBB2) and basal-like molecular subtype. In addition we identified a novel loss on 9p associated with the normal-like molecular subtype. This loss might induce a normal-like phenotype through the loss of the INK4a/ARF locus located at 9p21. Furthermore, KC-SMART identifies 18 recurrent aberrant regions in a new dataset of p53 conditional knock-out mouse model for breast cancer. These regions, combined with gene expression micro-array data, point to known cancer genes and novel candidate cancer genes.

**Variance Component Estimation for Class Discovery and Class Distinction in Microarray Data**

Greg Finak, McGill Center for Bioinformatics; Morag Park and Michael Hallett, McGill University
Often cancer expression profiling data sets are generated for exploratory purposes rather than to address a specific hypothesis. Consequently, it is difficult to design such experiments to optimally address all possible a posteriori questions. For example, since there are inherent difficulties associated with obtaining clinical samples and sensitive patient information, the amount of data available for analysis is limited and the presence of unobserved patient variables introduce potential confounding. Standard approaches to address confounding such as a balanced design are therefore not possible. When gene expression data is analyzed using classical approaches such as class distinction (identify all genes differentially expressed between two given classes), it may difficult to determine whether the genes identified are associated with the biological effect of interest, or whether they are associated with other unobserved, confounding effects. We address this issue using components of variance as a proxy for biological relevance. We reason that a set of genes with strong biological relevance to a particular class distinction should show regulatory control within the classes of interest. Consequently, such a set of genes should exhibit less variability within classes than between classes of interest. We estimate inter- and intra-class variance components from a random effects ANOVA using the method of moments, and develop a test statistic based on the ratio of Gamma distributed random variables to identify genes where inter-class variance component is greater than the intra-class variance component. This test, while similar in spirit to the classical ANOVA test statistic, differs in the underlying null hypothesis. Specifically, classical ANOVA tests whether the inter-class variance component is strictly greater than zero. Consequently, our test statistic is more stringent. We show simulation results to justify our modeling assumptions and compare our method against classical differential expression methods on real breast cancer data. We show, using HER2 positive vs. HER2 negative breast tumors, that our methodology identifies smaller sets of biologically relevant genes than classical tests of differential expression. This method may be added to existing analysis pipelines as an additional filtering step, reducing the number of candidate genes under consideration during the exploratory phase of a microarray study.

Applications of Custom, High Resolution DNA Microarrays to the Discovery and Characterization of Copy Number Variation in the Human Genome

Steve Laderman, Agilent Technologies; Nick Sampas, Amir Ben-Dor, Anya Tsalenko, Alicia Scheffer-Wong, Peter Tsang, Israel Steinfeld, Alice Yamada, Zohar Yakhini, Laurakay Bruhn (all Agilent)

Genome-wide copy number analyses using DNA microarray comparative genomics hybridization (aCGH) have revealed that submicroscopic insertions and deletions are a prominent mode of DNA sequence variation amongst apparently healthy individuals. The nucleotide percentage of the genome varying in this way is greater than and subject to higher mutation rates than variations in single nucleotide polymorphisms (SNPs). The overlap of CNVs with genes, untranslated RNA, and regulatory regions suggests that CNVs may play a broad role in determining the diversity of human traits, including the susceptibility to cancer and other diseases. The discovery and characterization of CNVs, however, is far from complete. Determinations of the sizes, positions, structure, allelic variation, and population frequency of CNVs are still underway while determinations of phenotypic consequence have scarcely begun. These goals are motivating the use and the further development of aCGH assays and data analysis methods. Motivations for refinement of the methods include, for example, the observation that an important class of CNVs is comprised of sequences that are duplicated in the reference genome. Another motivation is that determinations of CNV genotypes and the analysis of multiallelic states will be necessary to fully understand the phenotypic consequences of CNVs.

To illustrate these themes and show how highly reproducible, sensitive, and specific microarray methods can reveal new CNVs, clarify the boundaries, structure, and sequence characteristics of known CNV regions (CNVRs), and uncover how CNVs vary across populations, selected highlights from three studies will be reviewed. In one, DNA from thirty individuals from the HapMap collection was profiled using two custom 244K feature arrays with probe selection focused on previously reported CNVRs. The variations observed within the thirty HapMap samples overlapped the reported variations with greater than 90% concordance, but many regions were found to be significantly smaller. In another study, DNA from eight individuals from the HapMap collection were examined with two custom 244K arrays focused on intervals identified as variant with respect to the reference assembly by fosmid end-sequencing. Concordance of genotyping calls for thirty-five variant regions that had been previously characterized by sequencing and qPCR was 98%. Finally, applying a two-stage high resolution array approach to analyze fifty healthy Caucasian males from northern France, we found variants detected by more than one consecutive probe could be clustered into 1469 CNVRs, of which 721 are thought to be novel. These 721 novel regions contain 367 genes, 150 of which are represented in the Online Mendelian Inheritance in Man (OMIM) database. A further 6089 putative variants were detected by single probes: 48% of these were observed in more than one individual and 2662 lie outside previously reported CNVRs.

Prognosis of Prostate Cancer Relapse Using Gene Expression Profiling of FFPE Samples

Jean Lozach, Illumina, Inc.

Abstract TBA

Nexus CGH: A New Approach for Integrated Copy Number Analysis and Visualization

Soheil Shams, BioDiscovery

The use of genomic copy number data in cancer research is growing rapidly. It is well accepted that by effective integration of copy number data with a-priori data on genes, such as Gene Ontology data as well as experimental gene expression measurements, we can better identify genomic “hot spots” and prioritize genes as biomarkers as well as therapeutic targets. In this presentation, we will describe a new software tool developed by BioDiscovery, called Nexus CGH, which aims to provide a unique user interface
suitable for use by research scientist that can enable integration of data across multiple platforms and data modalities to quickly identify areas of genomic change. The software provides methods for identification of minimal areas of overlap in copy number data as well as statistical significance tests for copy number change frequency and multi-class comparisons. A live presentation will demonstrate the utility of this application on analysis of large scale public data.

**Hormone Regulated Expression and Splicing Using Exon and Junction Arrays**

Alexander Brodsky, Charles Lawrence, Luis Carvalho, Richard Park, Nicola Neretti and Sara Hillenmeyer, Brown University

We are using exon and junction arrays to probe how hormones such as estrogen and vitamin D make cell growth and death decisions. Our experiments focus on the early hormone regulated responses, where enrichment of specific motifs can be identified. We observe evidence of an alternative splicing program complementing the observed mRNA level changes. In the exon arrays, selection of significantly changing exons without transcript level changes leads to the identification of exon level changes in hundreds of genes with many being subject to alternative splicing. We also observe unique time behaviors for each hormone identified by hierarchical clustering. In order to analyze these arrays, we have developed new algorithms along with database and visualization tools. We find that these high density arrays are providing novel insights into the regulation of the transcriptome. In sum, we are uncovering new gene expression programs controlled by estrogen and vitamin D in cancer cells.

**Alternative Splicing in Breast Cancer Cell Lines: Whole-Genome Detection of Cell Specific Isoforms Using Novel Method FIRMA**

Anna Lapuk, Miki Yamamoto, John Conboy, Lawrence Berkeley National Laboratory; Ken Simpson, WEHI (Australia); Elizabeth Purdom, Terry Speed, Hovig Bayandorian, UC Berkeley; Rich Neve, LBNL; Tyson A. Clark and John E. Blume, Affymetrix; and Joe W. Gray, LBNL

Alternative splicing (AS) is a major mechanism of gene expression regulation. Up to 75% of all human genes undergo AS providing rich source for proteomic diversity. Splicing changes have been implicated in many diseases including cancer. Expression of splice isoforms with different functional properties, sometimes antagonistic, has been reported for many cancer genes, including CD44, p53, MDM2, BCL2, VEGF, FGFR1 and others. Differential expression of splice isoforms has been associated with tumor progression and has been used to improve tumor classification. Down regulation of cancer-specific splice forms can be achieved by RNA interference and cancer-specific protein isoforms can be targeted with antibodies. Thus, the knowledge of cell specific splice isoforms provides a potentially powerful resource for cancer diagnosis, prognosis, prediction and treatment.

The interrogation of splicing on the whole-genome scale is feasible using exon level expression data from exon and tiling arrays. However, the complexity of the array design and of the splicing phenomenon itself represents a significant challenge for detection of splicing events. Using novel method FIRMA (Finding Isoforms with RMA, in publication) in combination with Affymetrix Splicing Index we have developed an approach for effective detection of splicing changes in a complex background. Here we present a global analysis of splice variants in breast cancer by profiling breast cancer cell lines using Affymetrix Human Exon 1.0 ST arrays. More than a thousand exons emerged as strong candidates for alternative splicing. These included well-known genes (CD44, TPM1, ENAH, SH2B) and genes never reported to be alternatively spliced. Approximately half of the detected exons are novel, not included in the current gene structure. Splicing pattern has exhibited cell subtype specificity with exons preferentially included in luminal, basal A or basal B cells, most aggressive stem cell like subtype. A number of splicing events are specific to many cancer cells but not normal mammary epithelial cells. Splice patterns detected in this study will enable better understanding of breast cancer biology and can be used for development of effective new cancer biomarkers and therapeutic targets.

**Nova Regulated Aberrant Alternative Splicing in Cancer Tissues**

Bahar Taneri, Eastern Mediterranean University; and Terry Gaasterland, University of California, San Diego

Tissue-specific, regulated alternative splicing is critical to the maintenance and development of healthy cells. Aberrant alternative splicing is extensive in cancer tissues. Nova is a neuronal splicing regulator that governs the expression of neuron-specific forms of its target RNA transcripts. Nova targets a large number of pre-mRNA transcripts and controls the neuron-specific inclusion of some exons and exclusion of others. Nova autoregulates its own alternative splicing. The pathological expression of neuron-specific isoforms of Nova and its targets in lung and other cancer cells leads to an autoimmune response resulting in paraneoplastic neurologic disorder in some cancer patients. Correct expression of Nova is necessary for proper neuron development. Identification of misexpressed Nova targets in cancer cells may lead to insights into the cancer process. Here, we introduce a computational tool, termed as GEMhunter (Gene Expression Motif Hunter) and use it to search for Nova targets. With GEMhunter, we identified novel Nova targets based on clusters of weighted RNA binding motifs. GEMhunter applies normalized weight matrices for short motifs to nucleic acid sequences, identifies clusters of statistically significant matches, and generates a visualization of binding motif clusters and their statistical significance within sequences of interest. RNA binding motifs tend to be short. Nova binds to 4 nucleotides, two of which are degenerate: YCAY. When Nova binds in the intron near an exon, it enhances exon inclusion; binding within an exon promotes exclusion. Nova binding sites tend to co-occur near one another and even overlap, as in YCAYCAY or YCANYYCAY. In this study, GEMhunter was applied to the sequences of all cassette exons in all genes with variant isoforms in human, mouse and rat to screen for exonic clusters of Nova binding sites. Variant genes and their cassette exons were established by mapping expressed sequences
from healthy and cancer tissue libraries to genomes for each organism. GEMhunter identified 461 human, 136 mouse and 36 rat genes that contain one or more cassette exons predicted to be excluded by Nova. Functional analysis and cancer-specific expression analysis will further characterize the role of these genes in cancer cells. A list of Nova cassette exon targets is available via http://genomes.ucsd.edu/~bahar/GEMhunter/nova-targets.html.

Identification of Tumor-Specific Mutations in Coding Microsatellite DNAs by Using EST Data

Hojoon Lee and Stephen Johnston, Arizona State University

The elucidation of genetic alterations in cancer provides us with the opportunity to define tumor-specific antigens for cancer treatment. Microsatellite (MS) DNA, simple tandem repeat, is one of the main sources of genetic alterations in cancer because of extensive polymorphism and frequent occurrence in the human genome. Coding MS DNAs are potential sources to generate immunogenic tumor-specific antigens by frameshift mutation due to their propensity for insertion-deletion (Indels) mutations. We developed an algorithm to detect the tumor-specific Indels in MS DNAs in genes using the vast amount of expressed sequence tag (EST) data. We found that the mutation patterns of MS from our analysis of ESTs are quite consistent with the known characteristics of MS evolution, for example the preference of expansion to contraction, length-dependent mutability and variation by individual locus. Furthermore, the direct comparison of the frequency of Indels in the coding MS DNAs between cancer ESTs and normal ESTs indicated new target genes for instability in human cancers. This study has potentially generated new variants as cancer vaccine antigens.