The conference was of high quality and the keynote talks were superb. I will focus here on notable talks that are relevant to some group members. I recommend that everyone go through the conference book (we have hard copy and an electronic version) to get the broad pictures and catch up on the state of the art in other bioinfo topics. I'll leave NGS related topics for Roye's report.

- **On the Representation of de Bruijn Graphs**

The authors describe a navigational data structure that allows navigating a de Bruijn graph (going from a node to its neighbors) but not all functionality of previous data structures (e.g. it cannot tell if a node is in the graph). They provide lower bounds for the size of such data structure and implement one that his very compact – they reduce the size by a factor of about 2 compared to the best data structure of this type. Of theoretical and practical value.

**Relevant to Roye, Yaron.**

- **High-throughput genome scaffolding from in vivo DNA interaction frequency (HIGHLIGHT)**
  Noam Kaplan and Job Dekker.

The authors show that they can use contact maps of the genome obtained from HiC experiments in order to "close holes" in the genome and reconstruct the order of contigs by using the fact that contact strength along the chromosome goes down with the linear distance. Most genomes assembled today are in fact thousands of contigs and even in the human genomes there are contigs (or scaffolds) that are not placed. This proof of concept shows a very elegant way to use the HiC data to improve assemblies and close gaps.

**Relevant to Idan.**

- **An Exact Algorithm to Compute the DCJ Distance for Genomes with Duplicate Genes**
  M. Shao, Y. Lin, and B. Moret

An ILP formulation to the problem of finding a labeling of the different copies of each gene in a genome with duplicated genes so that the number of cycles is maximized. A cute trick is used to identify nodes/edges that are on the same cycle are identified. Outperforms previous method on human mouse rat comparisons (but the authors have to throw away all genes for which homolog detection is unsure (about 20%))

**Relevant to Ron.**
• An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding
  An interesting model for representing binding probabilities of transcription factors using ChIP-seq data from multiple experiments. This seems like a powerful method. Well worth the read (paper published in PLoS-CB)
  Relevant to Yaron.

• Learning Protein-DNA Interaction Landscapes by Integrating Experimental Data through Computational Models
  J. Zhong, T. Wasson, and A.J. Hartemink
  The authors develop an interesting approach to predict TF binding. They integrate ChIP-seq, DNAse sensitivity and MNase sensitivity experiments as well as prior information on PWMs. Their model is thermodynamic and gives strong results. Seems like an excellent work.
  Relevant to Yaron.

• Network based stratification of tumor mutations (Keynote)
  Trey Ideker
  Ideker gave a brilliant talk on two topics related to the challenge of using networks to translate genotype to phenotype. One (Hofree et al, Nature Methods 2013) uses the network and lists of mutated genes in many individuals in order to identify the disease causing genes. Involves "network smoothing" of the mutations and NMF clustering. None of the techniques is new but the idea is original and the results are impressive. The other paper uses the network to build a GO-like hierarchy, based on the insight that the "flat" network is hiding more complex structure of biology. Some parts of the GO hierarchy are rediscovered based on network info, and other parts possibly reveal new biology. The paper was published in Dutkowski et al. Nature Biotechnology 2013. Ideker's group has since created the NEXO website and web tool to support this new hierarchy.
  Relevant to Dvir (part 1) and Didi (both).

• Network deconvolution as a general method to distinguish direct dependencies in networks. (HIGHLIGHT)
  Soheil Feizi, Daniel Marbach, Muriel Medard and Manolis Kellis.
  This is the paper that generated the scandal between Pachter and Kellis. Irrespective of the drama, it is an interesting algebraic method to "clean" networks by identifying indirect relations that are caused by transitivity of correlation. A big part of the lecture was dedicated to refute the harsh allegations of Pachter.
  Relevant to Didi (for the network aspect) and all (since the whole community discusses it)
• **Finding Associations among Histone Modifications Using Sparse Partial Correlation Networks (HIGHLIGHT)**  
  Julia Lasserre, Ho-Ryun Chung and Martin Vingron.

An interesting methodology related to the previous paper (Kellis) using partial correlation to clean a network.

**Relevant to Didi**

• **Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data**  
  B.J. Raphael and F. Vandin

An interesting model and a new computation problem describing the progression of cancer as a sequence of steps where each step involves a mutation in one and only one gene out of a set of genes, and each step depends on all the previous steps being fulfilled. The sets correspond to pathways and the steps correspond to the progression of cancer. The model finds both the sets and the progression under a model allowing errors, using ILP.

**Relevant to Dvir.**

• **CSAX: Characterizing Systematic Anomalies in eXpression Data**  

A method to identify anomalies in expression data – given a training set of normal samples develop a way to identify if a new profile is abnormal. An interesting technique (based on a previously published paper of the authors) enhanced using GSEA. For the analysis the authors created a compendium of 28 human expression datasets that is made available on the web.

**Relevant to Kobi (?) Didi & Tom (the database).**

• **AptaCluster – A Method to Cluster HT-SELEX Aptamer Pools and Lessons from Its Application**  
  Jan Hoinka, Alexey Berezhnoy, Zuben E. Sauna, Eli Gilboa, and Teresa M. Przytycka

Using HT-SELEX for enrichment of aptamers binding to a particular target (protein/small molecule). Techniques are partially common with HTS of TFS. The authors developed a visualization tools AptaGUI that could be useful for us.

**Relevant to Yaron.**
To supplement Ron’s report, I’ve added my selections from both RECOMB and RECOMB-Seq. Some of our selections overlap; in those instances I’ve tried to highlight what I found novel, interesting, or potentially useful.

**RECOMB-Seq**

1) Near-optimal Assembly for Shotgun Sequencing with Noisy Reads - K. Lam, A. Khalak, D. Tse -

As opposed to modeling assembly as a combinatorial optimization problem imposed by reads of a certain technology, this work aims to characterize the length and coverage levels for which genome finishing is possible using the standard approaches - overlap layout consensus and de Bruijn graph traversal. This work extends one from last year’s RECOMB-Seq which assumed ideal, error free reads, to show how errors shift the information requirements.

2) Ensemble Analysis of Adaptive Compressed Genome Sequencing Strategies - Z. Taghavi -

The topic of this work is an ambitious question: how to identify every distinct genome in a heterogeneous microbial sample by sequencing and assembling. This topic has been recently investigated by Hamidreza Chitsaz’s group at Wayne State, and this work is an extension of earlier works. It describes a new simulation of a divide and conquer approach to attacking a currently theoretical (e.g., not experimentally implemented) technique for tackling the identification process.

3) Detecting Epigenetic Motifs in Low Coverage and Metagenomics Settings - N.D. Beckmann, S. Karri, G. Fang, A. Bashir -

PacBio reads are steadily gaining in popularity. They allow for read lengths long enough for genome finishing, and direct observation of epigenetic modifications at the single molecule level, but introduce significantly higher cost and error rates than Illumina reads do. This work attempts to identify methylated positions in bacterial samples via differences in kinetics observed during the sequencing procedure.

-Relevant to Yaron
RECOMB

1) IDBA - MTP: A hybrid metaTranscriptomic assembler based on protein information- H.C.M. Leung, S.M. Yui, F. Chin -
This work aims to solve the problem of frequent chimeras in metatranscriptomic assemblies. Their approach depends on mapping DNA to amino acid sequences. They build a de Bruijn graph (DBG) from read k-mers and align known proteins against paths in the DBG. Then, they look for shortest path aligning to proteins.

2) On the Representation of de Bruijn Graphs - R. Chikhi, A. Limasset, S. Jackman, J.T. Simpson, and P. Medvedev -
Ron also mentioned this paper. The first author developed Minia, the Bloom Filter (BF)-based assembler that I presented in our group meeting, and several of the other authors are also well known in the assembly niche. They show that by composing k-mers into maximal simple paths and adding an FM index for traversal you can perform assembly with much less memory than with BFs.

-relevant to Yaron

3) Traversing the k-mer landscape of NGS read datasets for quality score sparsification - Y.W. Yu, D. Yorukoglu, B. Berger –
This work is aimed at compressing quality (phred) scores of fastq files, which take up about twice the space of the read sequences but are harder to compress. The approach is based on k-mer counts, which avoids the burden of read alignment or BWT construction that competing methods depended on. They construct a dictionary of frequent k-mers among the reads, and then scan each read to identify k-mers in the read close (in terms of Hamming distance) to those in the dictionary. These k-mers are then used to smooth quality values for the positions covered by these k-mers. They show this allows for efficient lossy compression of the quality scores that surprisingly improves variant calling accuracy in their tests.

4) iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data (Highlight) - M.A. Mezlini, E. J. M. Smith, M. Fiume, O. Buske, G.L. Savich, S. Shah, S. Aparicio, D. Chiang, A. Goldenberg and M. Brudno -
The aim of this work is to combine identification of expressed isoforms in RNA-Seq with their quantification. Since the output of most quantifying tools is a set of relative abundances, the presence or absence of single transcripts may affect the
abundances of the rest. The authors start from nearly the set of all possible isoforms given known transcription start and end sites to achieve high sensitivity. They use regularized EM to reduce this set down to the transcripts actually expressed.

- relevant to Dvir

5) How the brain represents language and meaning - Tom Mitchell (Keynote talk)
Tom Mitchell discussed ongoing efforts to use fMRI to begin to understand how the brain encodes language. He described a process beginning with training classifiers to differentiate between word pairs, which was scaled up to observe similarities between different individuals’ representations of the same words, common words in different languages, comparisons between words and pictures representing the same objects, etc. He described applying their findings to predict brain regions activated in response to specific word queues, and provocative ideas for future works

- relevant to Didi and Adi

There have been several recent works aiming to use inferred genome blocks to identify spatial positions of ancestral origins (e.g., to place sequenced genomes on a map). The authors of this work take the reverse approach, aiming instead to leverage geographic information to improve haplotype phasing. Typically, phasing models assume that chromosomes are mosaics of blocks from past chromosomes in the population, and that in recombination block choices are made uniformly. Their contribution is to replace this uniformity assumption, based on the observation that nearby chromosomes are more likely to appear together.