Bioinformatics

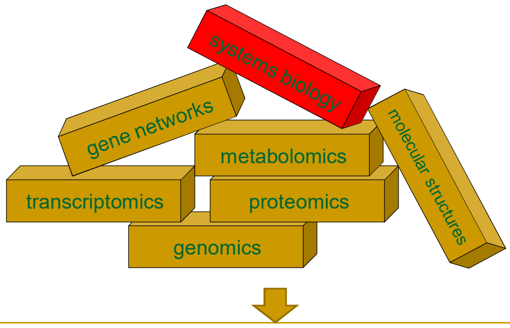
16. Review

陈铭（mchen@zju.edu.cn）
2016年1月11日

Credits

- Exam: 60%
- Class project (homepage): 30%
- Translation: 10%

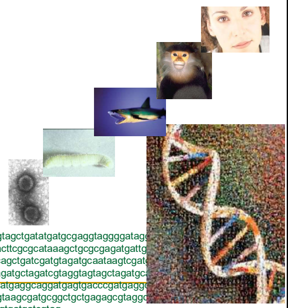
Full Review: Functional Bioinformatics



Synthetic Biology!

Computational Genomics

- Our genome encodes an enormous amount of information about our beings
  - our looks
  - our size
  - how our bodies work
  - ….
  - our health
  - our behaviors
  - … who we are?!



Computational Genomics

- Genetic parts-list encoded in a genome
  - genome sequence and variations
  - genomic structures
  - protein-coding genes
  - RNA-coding genes
  - pseudo genes
  - homologs/orthologs/paralogs
  - promoters/terminators
  - regulatory elements/binding motifs
  - transposable elements
  - …….

Computational Genomics

- To identify and characterize these elements, a large number of computational techniques have been developed and widely used in biological research
  - bio-sequence comparison
  - gene prediction
  - prediction of
    - orthologous genes
    - prediction of promoters
    - transcription factor binding motifs
    - operons
    - genome rearrangement
    - simple and complex repeats (http://10.71.115.210/minmat/)
    - SNPs and haplotype analysis
  - …….

Computational Genomics

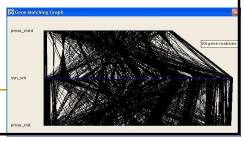
- suggesting biological functions of predicted genes, through homology search
- suggesting possible genes associated with a particular disease, and hence reducing the search space for relevant genes
- suggesting an organism's biology through genome comparison,
- suggesting component-candidate list and their possible interaction relationships in a biological pathway/network
- providing powerful tools for studies of biological evolution
  - sequence/genome comparison
  - phylogenetic profile analysis
- have played key roles in the human and other genome
  - genome assembly
  - protein-coding gene prediction
  - genome annotation

Challenges in Computational Genomics

- One challenge comes directly from the sheer amount of sequence data and the rate at which the data is being generated
  - Thousands of genomes have been (re)sequenced
  - Thousands of genomes are being sequenced
    - prokaryotic genomes / eukaryotic genomes

BIG DATA

- The amount of information potentially drivable through comparative genome analysis could be enormous knowing that functional elements are often conserved among "related" genomes
  - how to effectively derive them?!



Challenges in Computational Genomics

- Prediction of protein-coding genes still represents a challenging problem
  - accurate prediction of exon/intron boundaries
  - prediction of alternatively spliced gene forms
- Protein-coding genes account for ~3% of the human genome. What and where are the other "functional elements (ncRNAs?)" in the rest of the genome?
  - how to identify them?
  - how to (help to) predict their functions?

Challenges in Computational Genomics

- Identification of RNA-coding genes
  - what are the identifiable characteristics of RNA genes?
- Particularly, identification of small regulatory RNA
  - short interference RNAs (siRNA)
  - microRNA (miRNA)
  - Small RNA (smRNA)
- Identification of regulatory elements/binding sites
  - transcription regulatory binding sites
  - splice factor binding sites
  - other classes of regulatory elements?

Challenges in Computational Genomics

- Identification of other types of functional elements
  - transposons
  - ….
- Identification of genome variations – polymorphisms
  - identification of SNPs
  - prediction of haplotype blocks
- Recognition of genome structures
  - operons, regulons in microbes
  - genomic structures in eukaryotic genomes

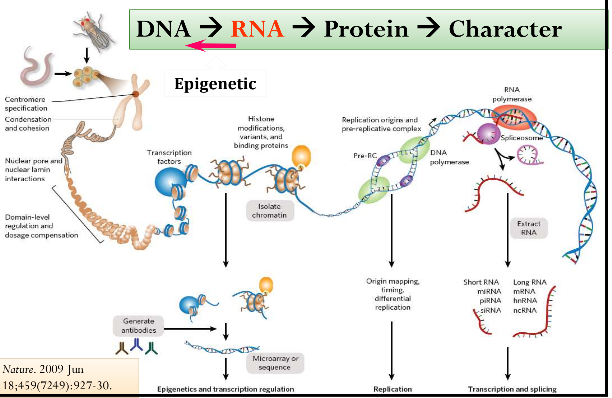
Challenges in Computational Genomics

- Genome is not a linear sequence; It is a 3D structure! 3D Genome
  - accurate identification and characterization of functional elements by looking at the genome as a 3D DNA structure

…. and many other outstanding challenges!

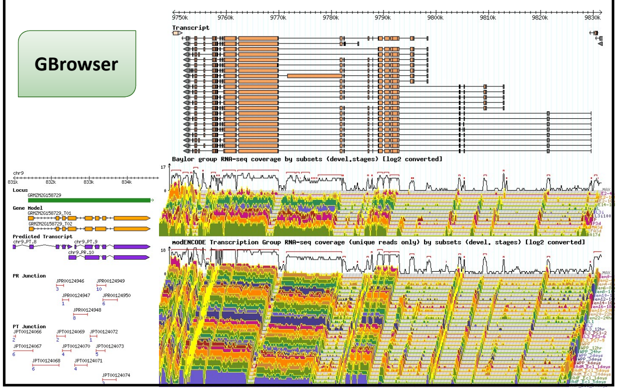
Sequence-based Functional Elements on Central Dogma

DNA → RNA → Protein → Character

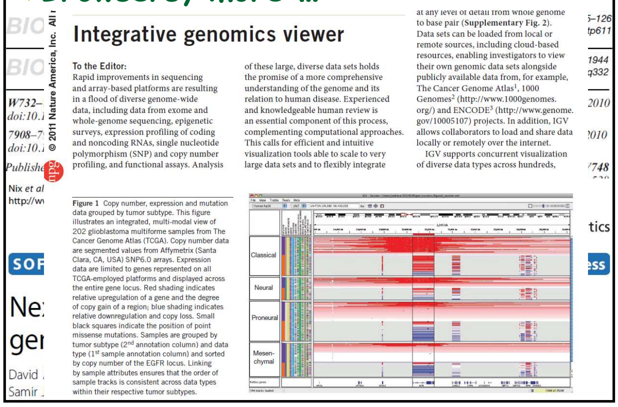


Nature. 2009 Jun 18;459(7249):927-30.

Web-based Browsers



Browsers, more …



Integrative genomics viewer

To the Editor:
Rapid improvements in sequencing and array-based platforms are resulting in a flood of diverse genome-wide data, including data from exome and whole-genome sequencing, epigenetic surveys, expression profiling of coding and noncoding RNAs, single nucleotide polymorphism (SNP) and copy number profiling, and functional assays. Analysis of these large, diverse data sets holds the promise of a more comprehensive understanding of the genome and its relation to human disease. Experienced and knowledgeable human review is an essential component of this process, complementing computational approaches. This calls for efficient and intuitive visualization tools able to scale to very large data sets and to flexibly integrate

at any level of detail from whole genome to base pair (**Supplementary Fig. 2**). Data sets can be loaded from local or remote sources, including cloud-based resources, enabling investigators to view their own genomic data sets alongside publicly available data from, for example, The Cancer Genome Atlas[1], 1000 Genomes[2] (http://www.1000genomes.org/) and ENCODE[3] (http://www.genome.gov/10005107) projects. In addition, IGV allows collaborators to load and share data locally or remotely over the internet.

IGV supports concurrent visualization of diverse data types across hundreds,

Figure 1 Copy number, expression and mutation data grouped by tumor subtype. This figure illustrates an integrated, multi-modal view of 202 glioblastoma multiforme samples from The Cancer Genome Atlas (TCGA). Copy number data are segmented values from Affymetrix (Santa Clara, CA, USA) SNP6.0 arrays. Expression data are limited to genes represented on all TCGA-employed platforms and displayed across the entire gene locus. Red shading indicates relative upregulation of a gene and the degree of copy gain of a region; blue shading indicates relative downregulation and copy loss. Small black squares indicate the position of point missense mutations. Samples are grouped by tumor subtype (2nd annotation column) and data type (1st sample annotation column) and sorted by copy number of the EGFR locus. Linking by sample attributes ensures that the order of sample tracks is consistent across data types within their respective tumor subtypes.

Applications of RNA-Seq

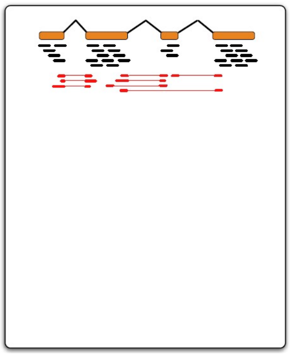
- Expression analysis
- Discovery of new transcribed regions → find new genes
- Discovery of alternative splicing → determine the transcriptional structure of genes
- Discovery of gene fusion
- Discovery of small ncRNAs/microRNA primary transcripts
- Discovery of SNPs/CNVs/RNA editing sites
- Small RNA transcriptome → smRNA-seq

Alternative splicing

- Expansion of transcriptome → proteome → phenotype.
  - Human: ~95% of multi-exon genes;
  - fly: ~61% of multi-exon genes or ~31 % all genes;
  - rice: ~33%~48% all genes;
  - Arabidopsis: ~42% intron-containing genes;
  - maize: ~56% of intron-containing genes;
  - nematode: ~25% all genes.
- Aberrant RNA splicing → diseases
- *cis*-splicing

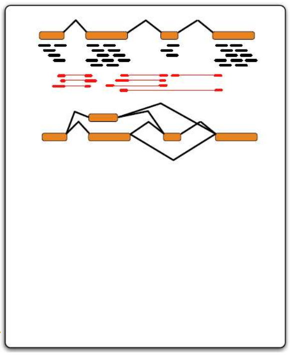
Infer alternative isoforms

- ✓Map reads
- ✓Build splicing graph using spliced reads
- ✓Generate transcripts from graph
- ✓explain coverage by sparse, weighted sum of transcripts



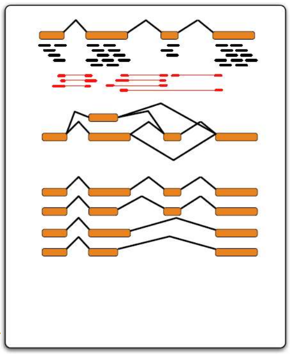
Infer alternative isoforms

- ✓Map reads
- ✓Build splicing graph using spliced reads
- ✓Generate transcripts from graph
- ✓explain coverage by sparse, weighted sum of transcripts



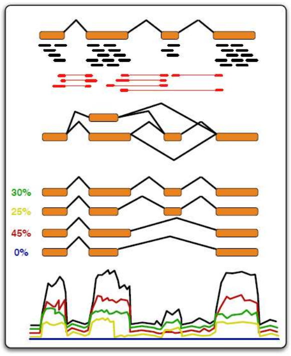
Infer alternative isoforms

- ✓Map reads
- ✓Build splicing graph using spliced reads
- ✓Generate transcripts from graph
- ✓explain coverage by sparse, weighted sum of transcripts



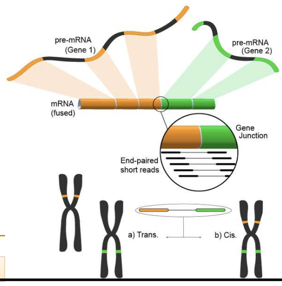
Infer alternative isoforms

- ✓Map reads
- ✓Build splicing graph using spliced reads
- ✓Generate transcripts from graph
- ✓explain coverage by sparse, weighted sum of transcripts



Gene fusion

- fusion genes → cancer → "cancer genes"
- resulting from chromosomal rearrangements in cancer
- *trans*-splicing



[Wikipedia]

Discovery of ncRNAs

- Small ncRNAs, Long ncRNAs (lncRNAs)
- New putative lncRNAs generally identified by RNA-seq → RNA immunoprecipitation followed by sequencing (RIP-seq), parallel analysis of RNA structure (PARS), fragmentation sequencing (Frag-seq) …

RNA-DNA Differences

- SNPs/CNVs
- RNA editing

Sciencexpress Research Article

Widespread RNA and DNA Sequence Differences in the Human Transcriptome

Mingyao Li,[1]* Isabel X. Wang,[8]* Yun Li,[6,7] Alan Bruzel,[8] Allison L. Richards,[4] Jonathan M. Toung,[5] Vivian G. Cheung[2,3,8]†

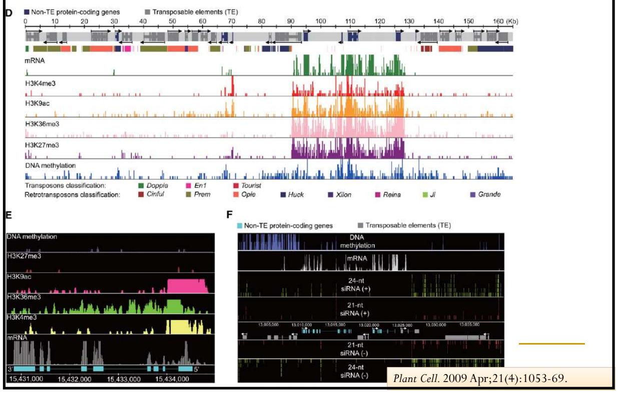
[1]Departments of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [2]Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [3]Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [4]Cell and Molecular Biology Graduate Program, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [5]Genomics and Computational Biology Graduate Program, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA. [6]Departments of Genetics, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA. [7]Department of Biostatistics, University of North Carolina School of Medicine, Chapel Hill, NC 27599, USA. [8]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA.

Science. 2011 May 19.

Integrated maps of the epigenome and transcriptome in Arabidopsis

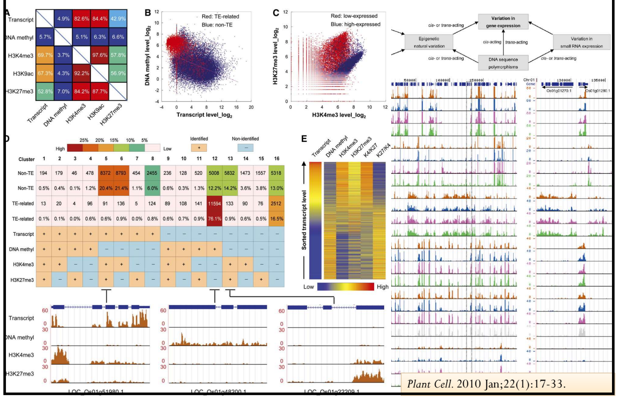
Cell. 2008 May 2;133:523-536.

Epigenetic and transcriptional maps in maize



Plant Cell. 2009 Apr;21(4):1053-69.

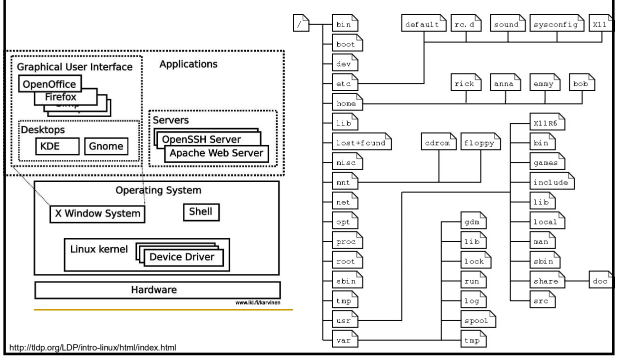
Epigenetic and transcriptional maps in rice



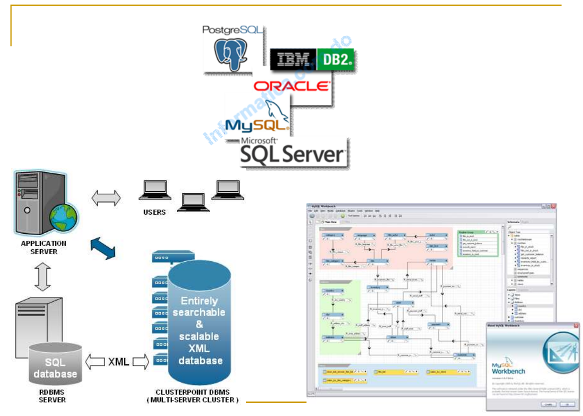
Plant Cell. 2010 Jan;22(1):17-33.

Biocomputing

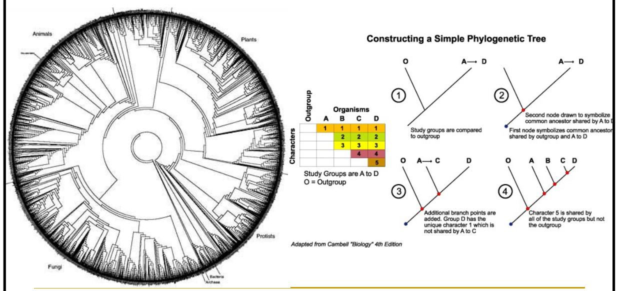
Linux system



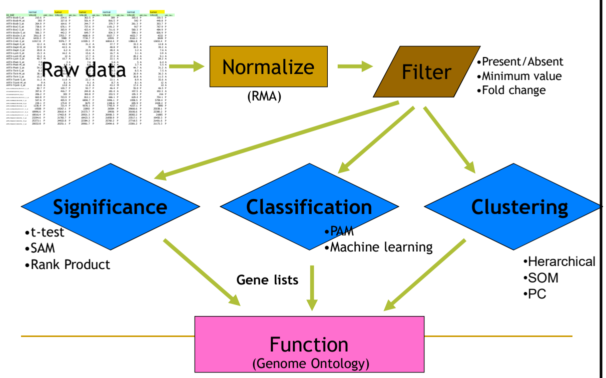
http://tldp.org/LDP/intro-linux/html/index.html



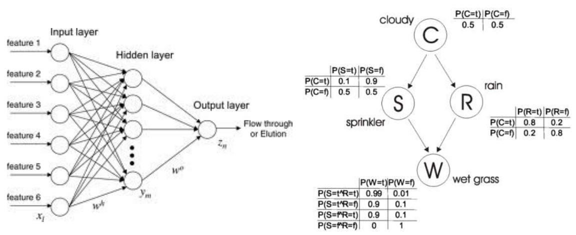
Phylogenetic tree



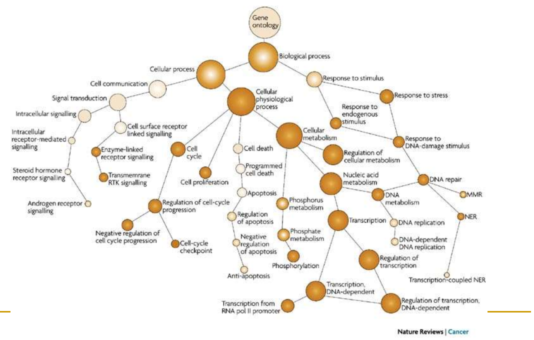
Streamlined Affy Analysis



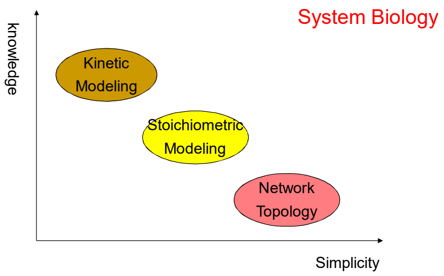
Computational biology

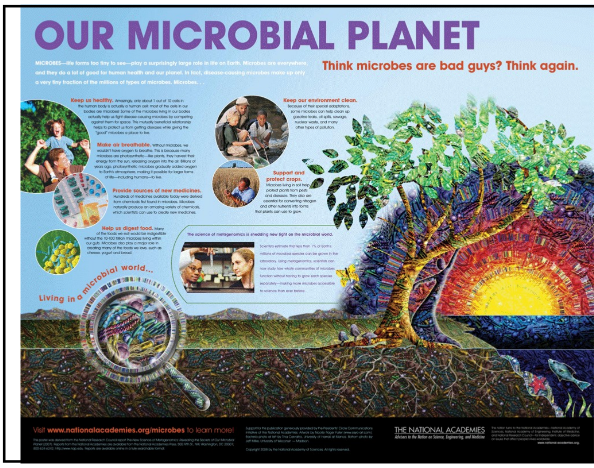


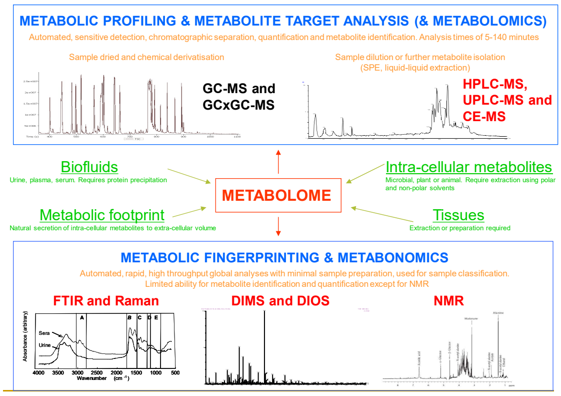
GO

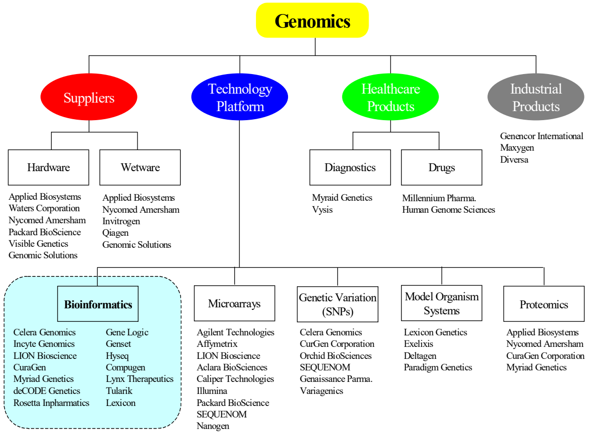


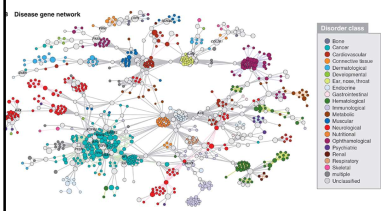
Biological network modeling

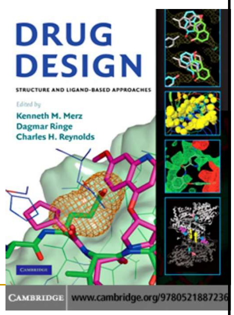


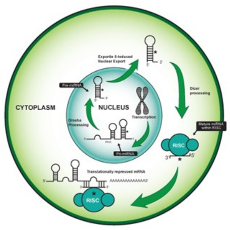


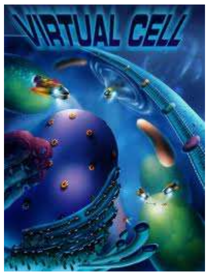












Open Questions

- From molecule motor to ecology mode

- Surface curve (apple, peach…)
- Cell movement
- Polygenetic tree
- Protein topology, Image recognition
- Network analysis, comparison, organization
- Virtual cell modeling and simulation