Computing Basics for Genomic Analysis

January 28, 2013

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Schedule

Start	End	Topics
9:00:00	9:30:00	Registration and Overview
9:30:00	10:30:00	Genomic Data Files
10:30:00	11:30:00	Unix Basics 1 - Server and commands
11:30:00	12:30:00	Unix Basics-2 – Shell scripting
12:30:00	13:30:00	Lunch
13:30:00	14:45:00	R basics
14:45:00	15:00:00	Break
15:00:00	16:30:00	R visualization and example
16:30:00	17:00:00	NGS applications

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

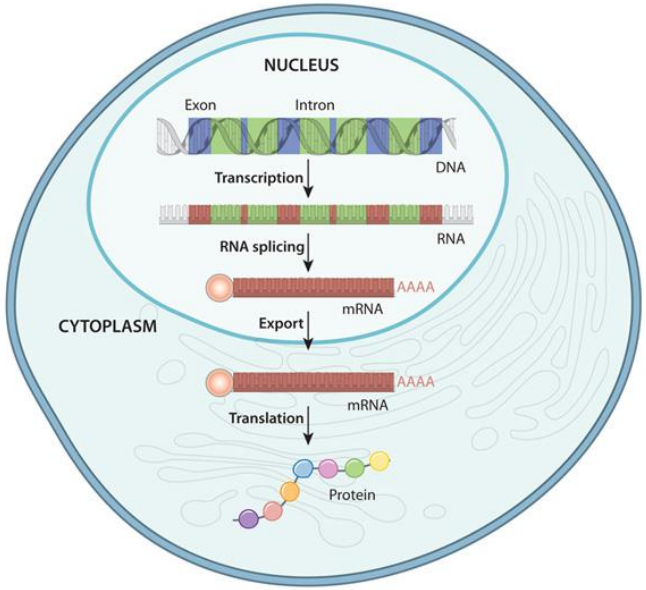
Computing Basics for Bioinformatics

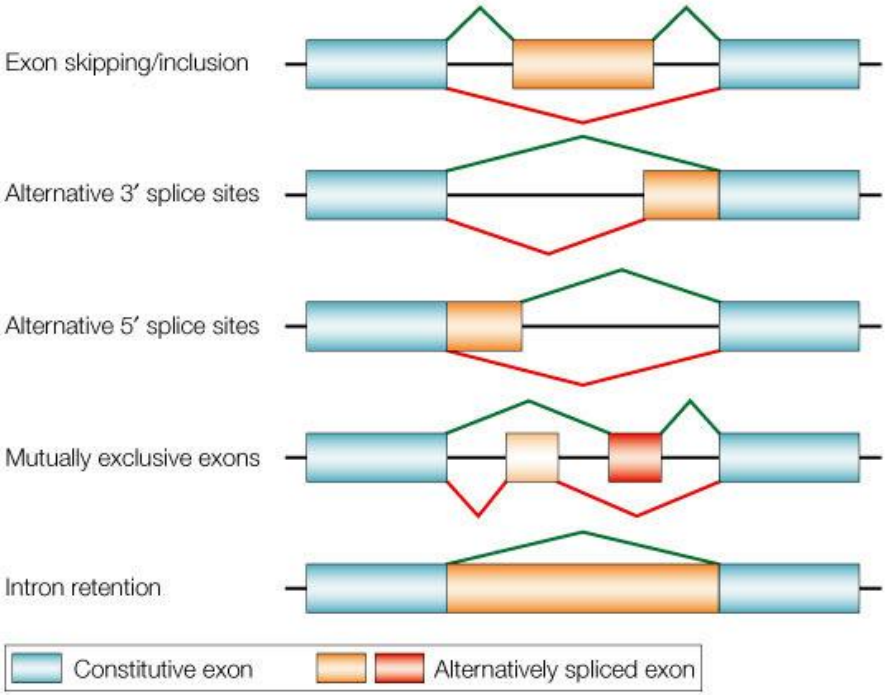
Session 1

Yaoyu Wang

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Central Dogma of Molecular Biology





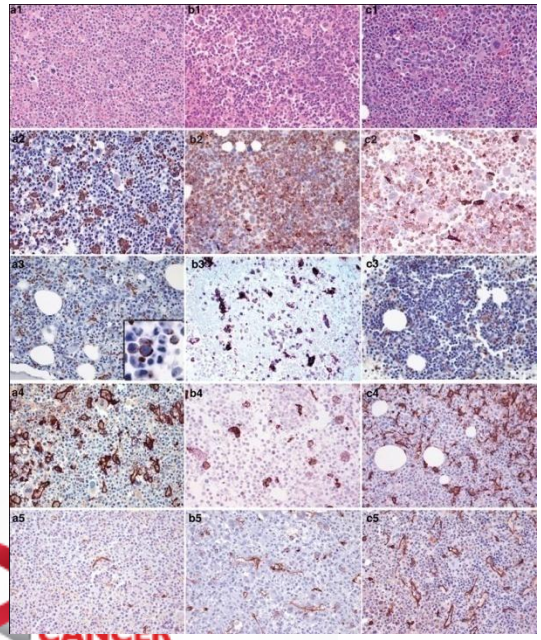
CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Cartegni et al, Nature Reviews Genetics, (2002)

What initial questions do we want to answer using genomic data?

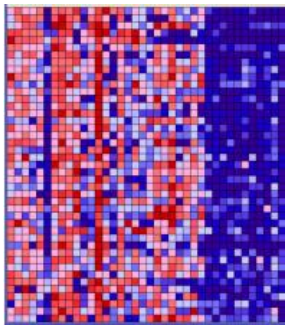
- Which genes are expressed?
- Which genes are differentially expressed among the groups?
- Do my favorite genes/pathways become up/down-regulated?
- Can we detect RNA isoforms? Novel ones?
- What are the genomic regions with copy number variations?
- Can we detect structural variants? SNPs, insertions, deletions, RNA-editing?
- What are the TF binding sites?

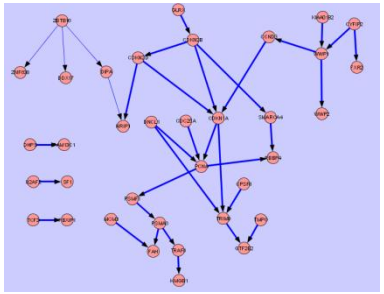
Genomics Research

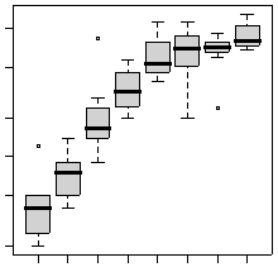


Microarray

NGS







CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Next-Generation Sequencing Workflow

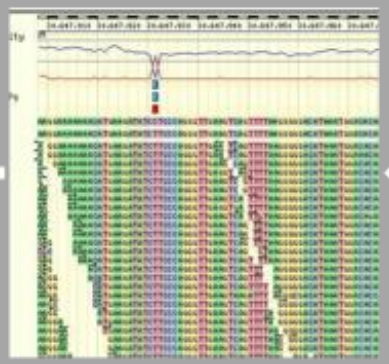
1. Sample Preparation

2a. Flow Cell

2b . Cluster Generation

3 . Sequencing & Imaging

4 . Initial Image Analyses

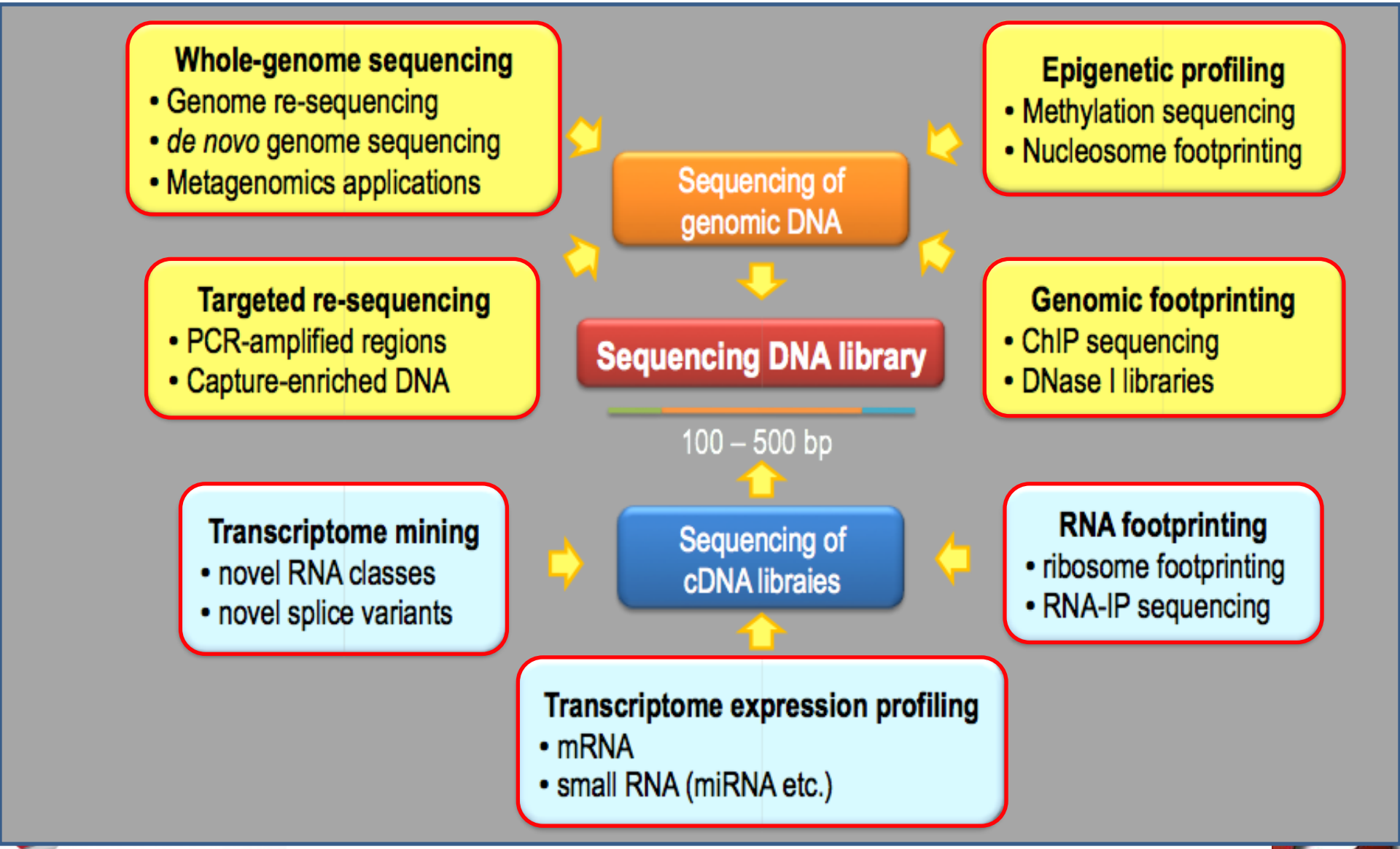


5 . Data Analysis

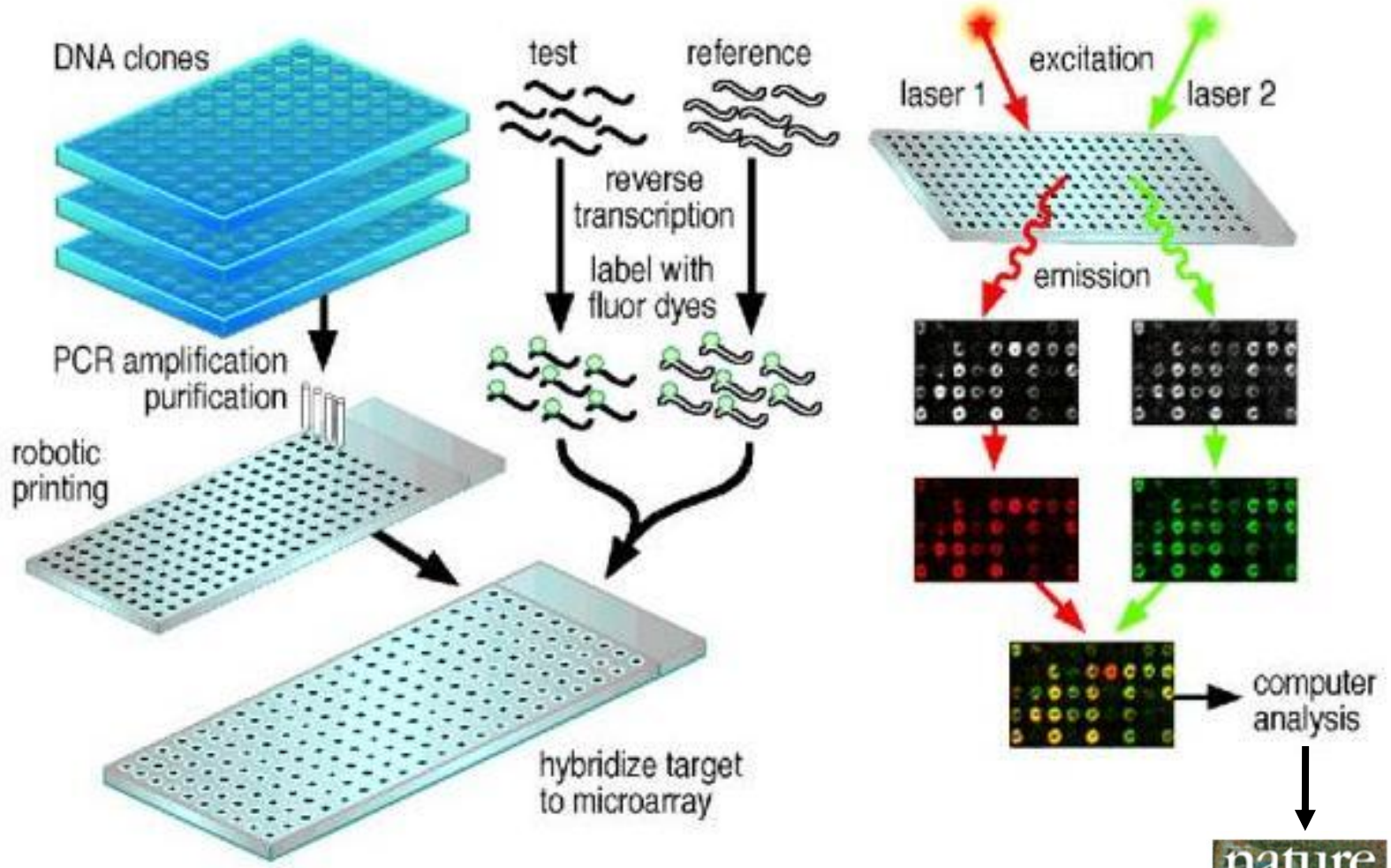
6 . Publication

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Applications of Next-Generation Sequencing

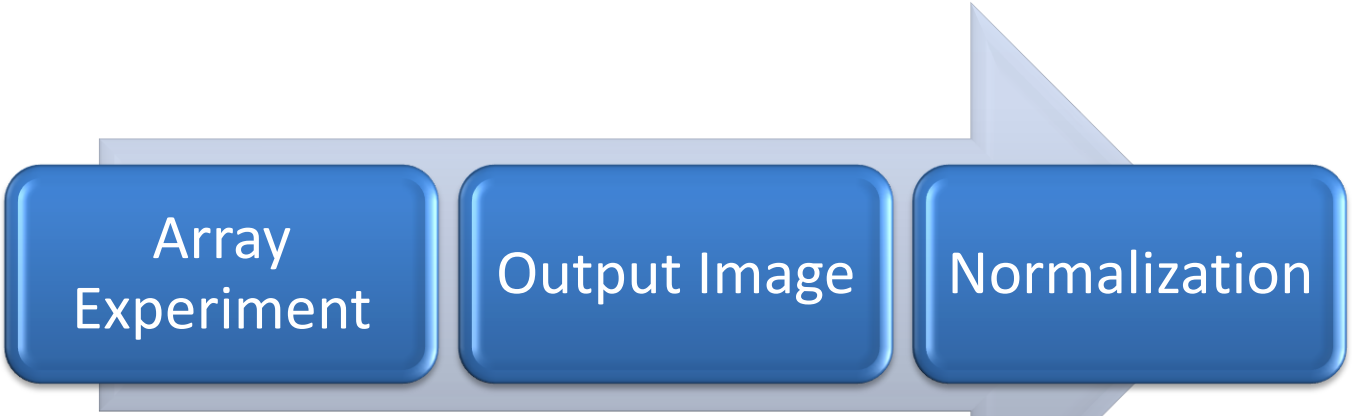


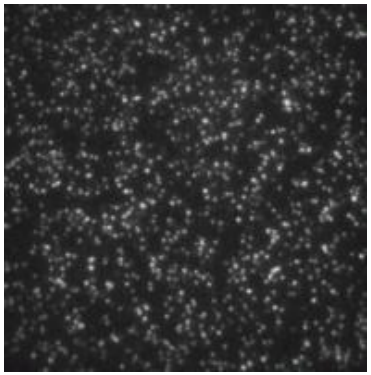
Microarray Experiment Flow



CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

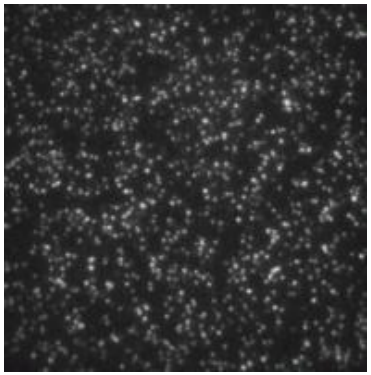
Data Generation Pipeline-NGS





CEL file

Data Generation Pipeline-Microarray



Fastq

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Data Analysis- Individual Sample Data is huge

Per human sample

Platform	Exome/aCGH	Transcriptome	ChIP
Micoarray	~115 MB	~13-40 MB	~60 MB
HiSeq Run	~67.5-135 GB	~33.75-67.5 GB	~13.5-20 GB

For 20 human sample

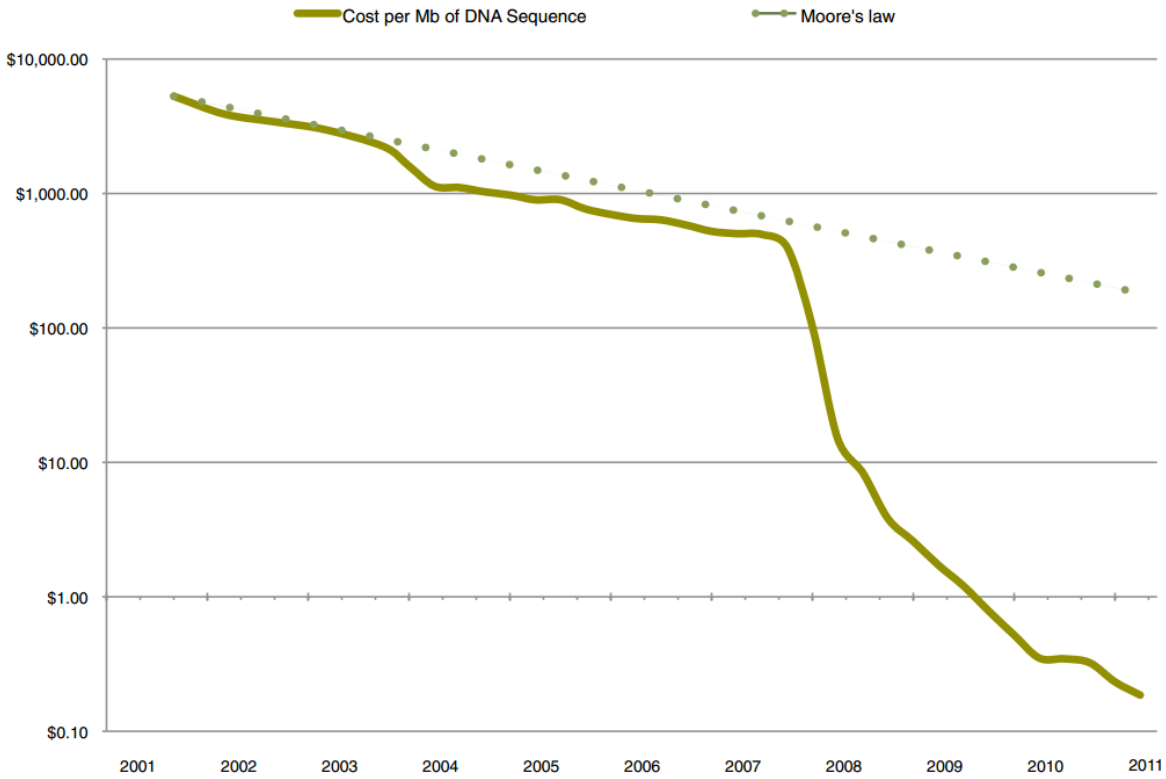
Platform	Exome/aCGH	Transcriptome	ChIP
Micoarray	~2 Gb	~260 MB	~1.2 MB
HiSeq Run	~1350 GB	~680 GB	~270 Gb

MacBook Pro

Price	$2,799
Processor	2.6GHz, quad-core
Memory	16 GB
Storage	768 GB

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Data Analysis-DNA Sequencing Cost is Rapidly Dropping



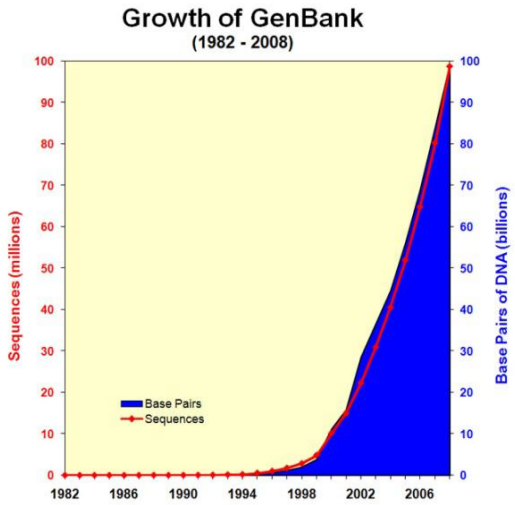


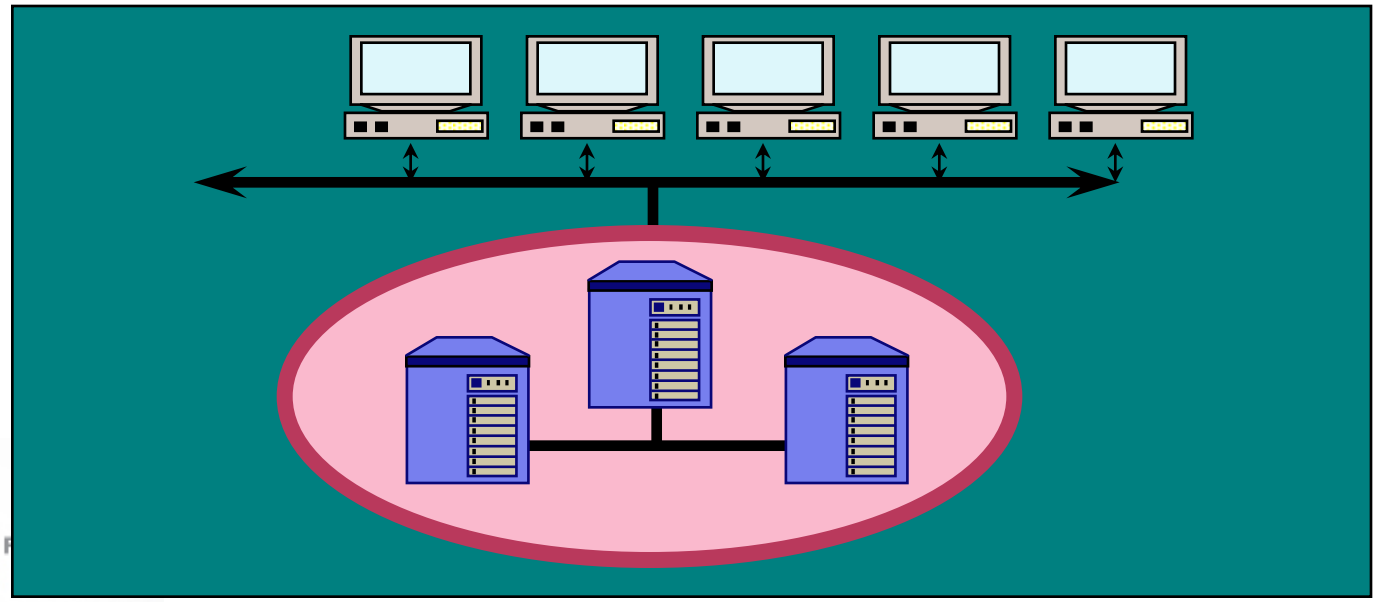
Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

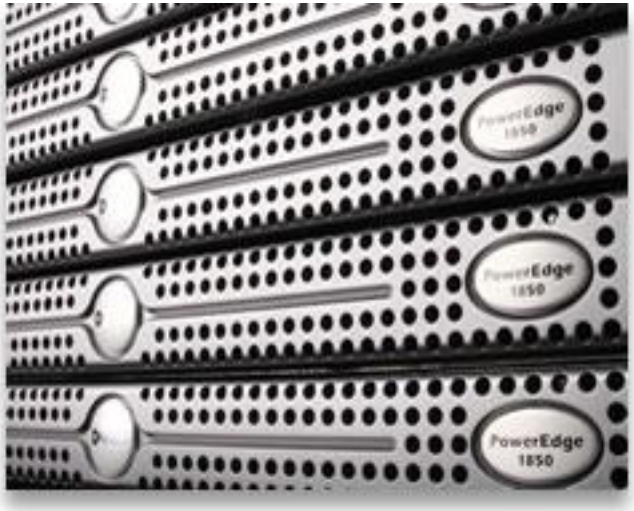
CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

The first genome took 10 yrs and $3 billion
Today, 1 week for $5000

Genomics Data analysis is a job for computer Cluster

- Group of independent systems that
  - Function as a single system
  - Appear to users as a single system
  - And are managed as a single system'
- Clusters are "virtual servers"





The cluster is a stack of "rack mount" servers a bit like pizza boxes

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE

Cluster Resources on Longwood Campus

- Partners Research Computing
  - http://rc.partners.org
- Research IT Group of HMS
  - http://ritg.med.harvard.edu/
- Research Computing at DFCI
  - http://research4.dfci.harvard.edu/
- Amazon Cloud Computing

CENTER FOR CANCER COMPUTATIONAL BIOLOGY
DANA-FARBER CANCER INSTITUTE