Computational Challenges in Storage, Analysis and Interpretation of Next-Generation Sequencing Data

Shouguo Gao Ph. D

Department of Physics and Comprehensive Diabetes Center

Next Generation Sequencing

Technologies that parallelize the sequencing process, producing thousands or millions of sequences (DNA/RNA) at once.

(Wikipedia)

Should be “current generation sequencing”

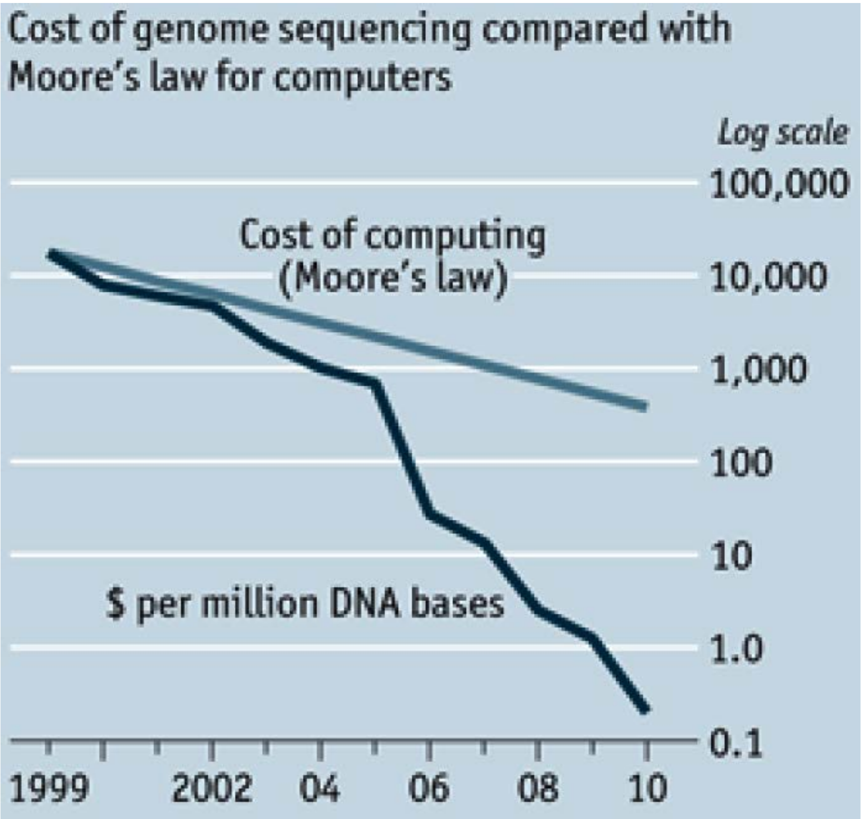
Low cost

High speed

Next Generation Sequencing

- DNA-seq (whole genome DNA or exons only)
- RNA-seq (RNA)
- Chip-seq (Transcriptional factors binding sites)
- DNase-seq, FAIRE-seq (histone modification, all transcriptional binding sites and other modifications)
- Methylation (DNA methylated sites)
- etc.

Next Generation Sequencing



http://www.economist.com/node/16349358

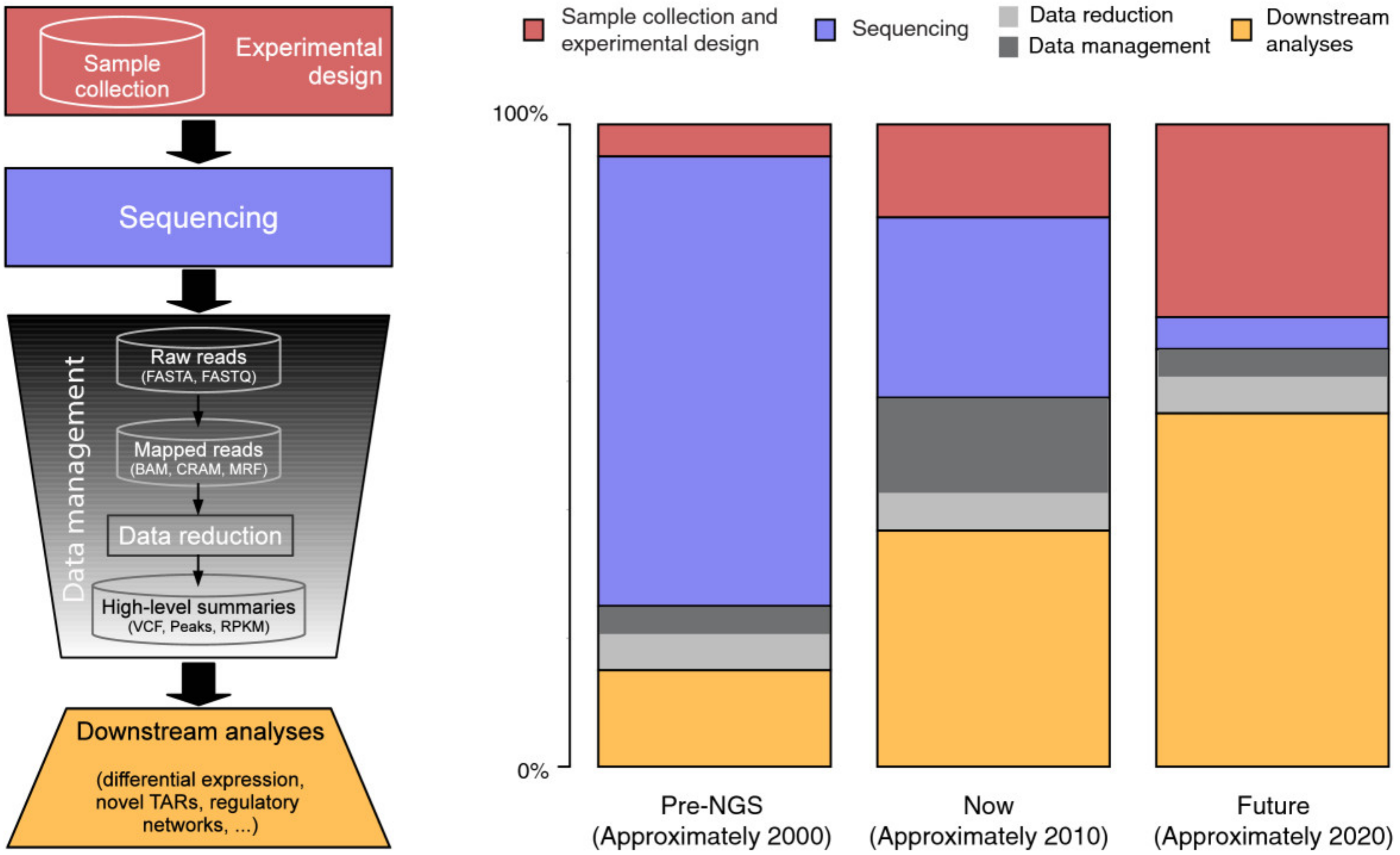
Roche/454 FLX Titanium
400-600 million reads/run
400bp avg. length

Illumina HiSeq 2000
Up to 6 billion PE
reads/run
35-100bp read length

SOLiD 4
1.4-2.4 billion PE reads/run
35-50bp read length

Next Generation Seqnecing

- cost of sequencing the first human genome was about **$3 billion**, and it took several international institutes, hundreds of researchers and **13 years** to complete (1990-2003).
- In 2007, James Watson’s genome completed for less than **$1million** in **several months** (Roche 454)
- By 2009 the cost for a whole-genome sequence dropped to **$100,000**
- By 2012 **$1,000 Genome in Two Hours** by Ion Torrent (still high error rate)



Snoner et al, Genome Biology 2011, 12:125

Storage Challenge

Table: Storage space for mouse stem cell sample

	Whole genome Sequencing	RNA-Seq
Sequencing Storage	~300GB (30x coverage)	~30GB
Downstream analysis	~310GB	~30GB

Garber et al, Nature Method 2011

The space for storing the data is increasing at a slower pace than sequencing throughput. We will have no enough storage space in future.

Storage Challenge

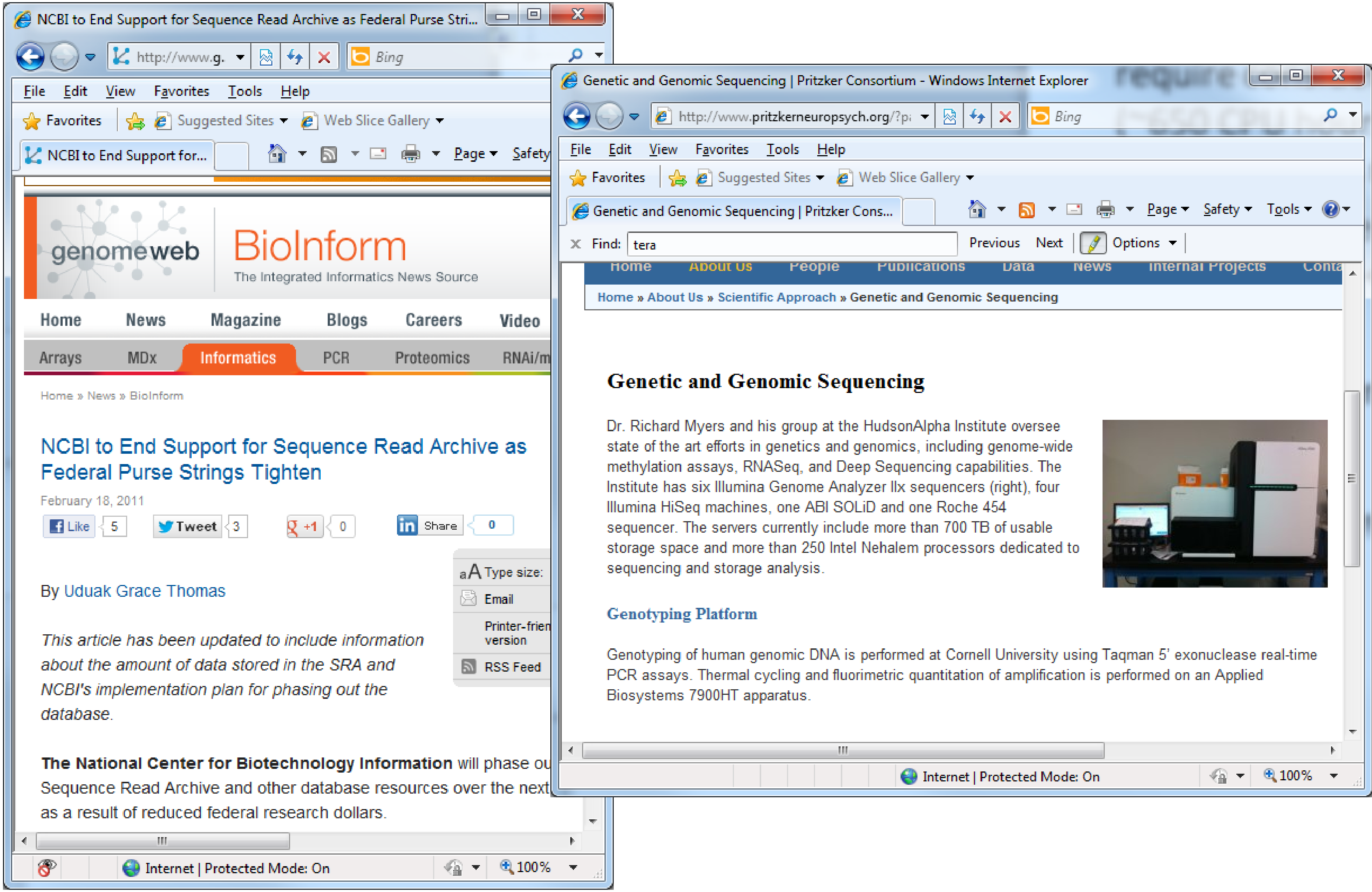
- Data transfer is another problem

Table: Time and cost of data transfer

	Whole genome Sequencing	RNA-Seq
Transfer (10MB/s)	12h	Few hours
cost	$40	$5

Garber et al, Nature Method 2011

local area network in HudsonAlpha can operate at speeds up to 10 GB/s



NCBI to End Support for Sequence Read Archive as Federal Purse Stri...
http://www.g.
Bing
File Edit View Favorites Tools Help
Favorites
Suggested Sites
Web Slice Gallery
NCBI to End Support for...
Page
Safety
genomeweb
BioInform
The Integrated Informatics News Source
Home News Magazine Blogs Careers Video
Arrays MDx Informatics PCR Proteomics RNAi/m
Home » News » BioInform

NCBI to End Support for Sequence Read Archive as Federal Purse Strings Tighten

February 18, 2011

Like 5
Tweet 3
+1 0
Share 0
aA Type size:
Email
Printer-frien version
RSS Feed

By Uduak Grace Thomas

This article has been updated to include information about the amount of data stored in the SRA and NCBI's implementation plan for phasing out the database.

The National Center for Biotechnology Information will phase ou
Sequence Read Archive and other database resources over the next
as a result of reduced federal research dollars.

Internet | Protected Mode: On
100%
Genetic and Genomic Sequencing | Pritzker Consortium - Windows Internet Explorer
http://www.pritzkerneuropsych.org/?p
Bing
File Edit View Favorites Tools Help
Favorites
Suggested Sites
Web Slice Gallery
Genetic and Genomic Sequencing | Pritzker Cons...
Page
Safety
Tools
Find: tera
Previous
Next
Options
Home About Us People Publications Data News Internal Projects Conta
Home » About Us » Scientific Approach » Genetic and Genomic Sequencing

Genetic and Genomic Sequencing

Dr. Richard Myers and his group at the HudsonAlpha Institute oversee state of the art efforts in genetics and genomics, including genome-wide methylation assays, RNASeq, and Deep Sequencing capabilities. The Institute has six Illumina Genome Analyzer IIx sequencers (right), four Illumina HiSeq machines, one ABI SOLiD and one Roche 454 sequencer. The servers currently include more than 700 TB of usable storage space and more than 250 Intel Nehalem processors dedicated to sequencing and storage analysis.

Genotyping Platform

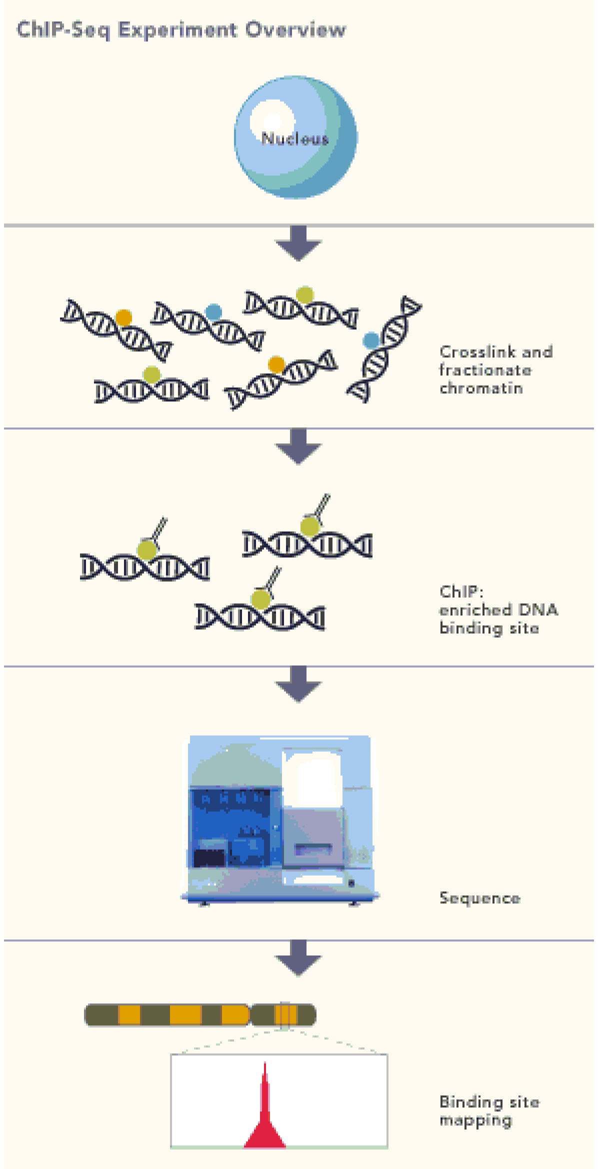
Genotyping of human genomic DNA is performed at Cornell University using Taqman 5' exonuclease real-time PCR assays. Thermal cycling and fluorimetric quantitation of amplification is performed on an Applied Biosystems 7900HT apparatus.

Computation challenge

Several examples

Chip Seq

- 60-90% DNA are not Transcript factor specific.
- I checked the codes of 6 open-source software. The assumptions hidden in the code are quite different and the results are not consistent very good. ~ 70% overlap in the identified peaks
- Good and widely accepted algorithm to lower the False Positive Rate is still unavailable.


ChIP-Seq Experiment Overview
Nucleus
Crosslink and fractionate chromatin
ChIP: enriched DNA binding site
Sequence
Binding site mapping

RNA seq mapping

- Reason of complexity

  No perfect gene annotation

  Sequencing error

  Short fragments

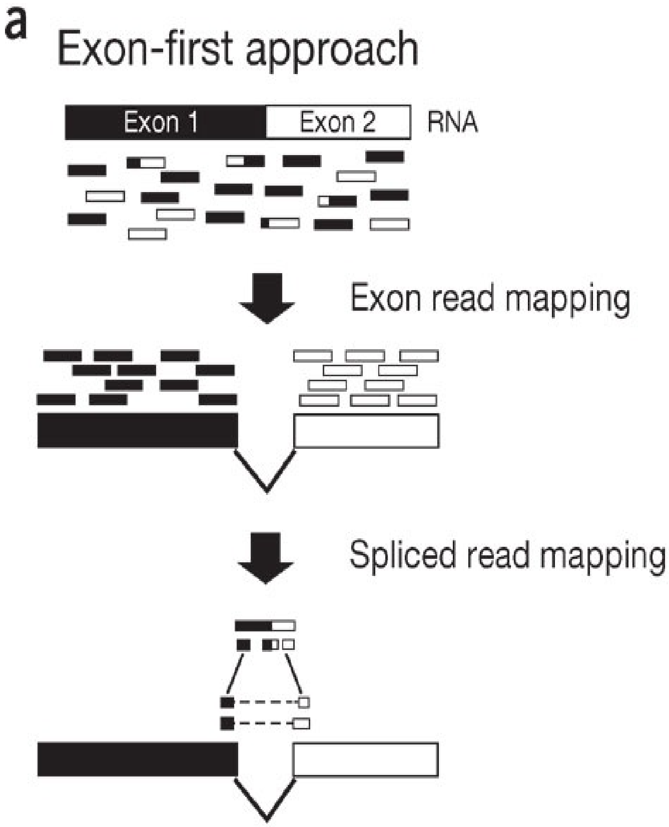
  Millions of reads

  Reads span exon-exon junctions

- Time for mapping to genome

  110 CPU hrs with good annotation

  1000 CPU hrs without good annotation



Garber, et al, Nature Method, May, 2011

Transcriptome Reconstruction

Defining a precise map of all transcripts and isoforms that are expressed in a particular sample.

More specifically, we need the assembly of all reads or read alignments into transcription units.

Transcriptome Reconstruction

- Without a good reference genome

  ~650 CPU hours and >16 GB of RAM

- With good reference genome

  ~4 CPU hours and ~4 GB RAM

Garber, et al, Nature Method,
May, 2011

Just for one mouse stem cell sample!!!

My experience

- cufflinks for RNA-seq data for Type 1 Diabetes

  5 samples and 5 controls

  14 days on our ZEN server

  (8 dual core processors, ~16GB memory)
- Our server is not powerful enough to deal with the NGS data from TCGA project (several hundreds samples).

Meta-genome sequencing

Genetic material recovered directly from environmental samples. The data contains fragmented data representing as many as 10,000 species.

The human gut microbiome gene catalog identified 3.3 million genes assembled from 567.7 GB of sequence data (Qin,J et al, Nature 2010, 464: 59–65).

Collecting, curating, and extracting useful biological information from datasets represent significant computational challenges for researchers.

Acknowledgments

- Dr. Xujing Wang’s group
- Department of Physics
- Comprehensive Diabetes
- NIDDK