Biology and Big Data

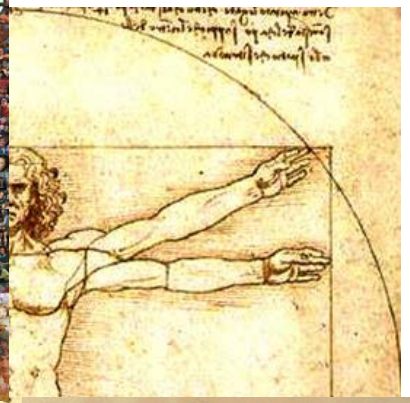
W. Brad Barbazuk

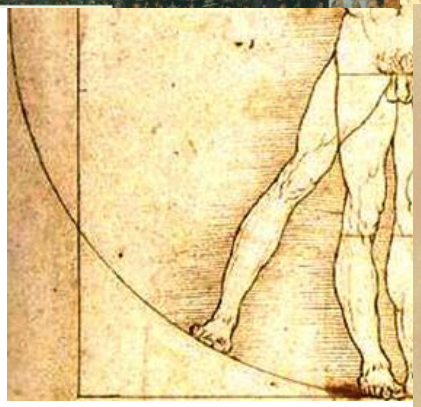
Department of Biology, and the
University of Florida Genetics Institute

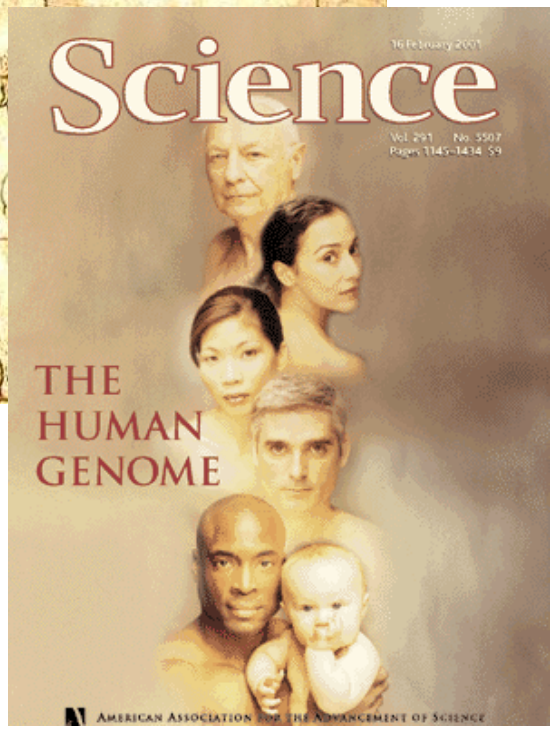
Genomics

- The Study of Genomes
  - Determine the entire DNA sequence of an organism.
  - Identify “features” within the Sequence.
  - Determine how/when genes are turned on/off.
  - Identify differences in DNA sequence between members of a population – correlate these with phenotype

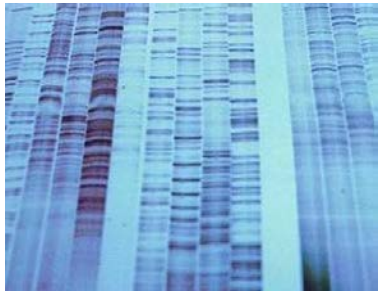
- In 2001, the sequence of the human genome was reported.
- 15 year project, $2.9B
- 21st century equivalent of putting a man on the moon.



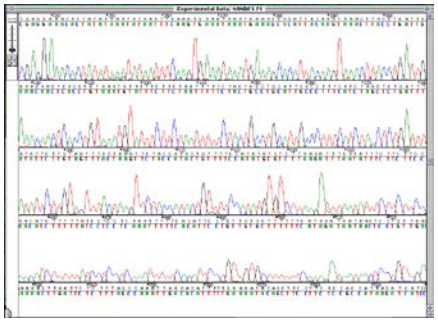




Primary sequence data



1980s – 200-300bp.

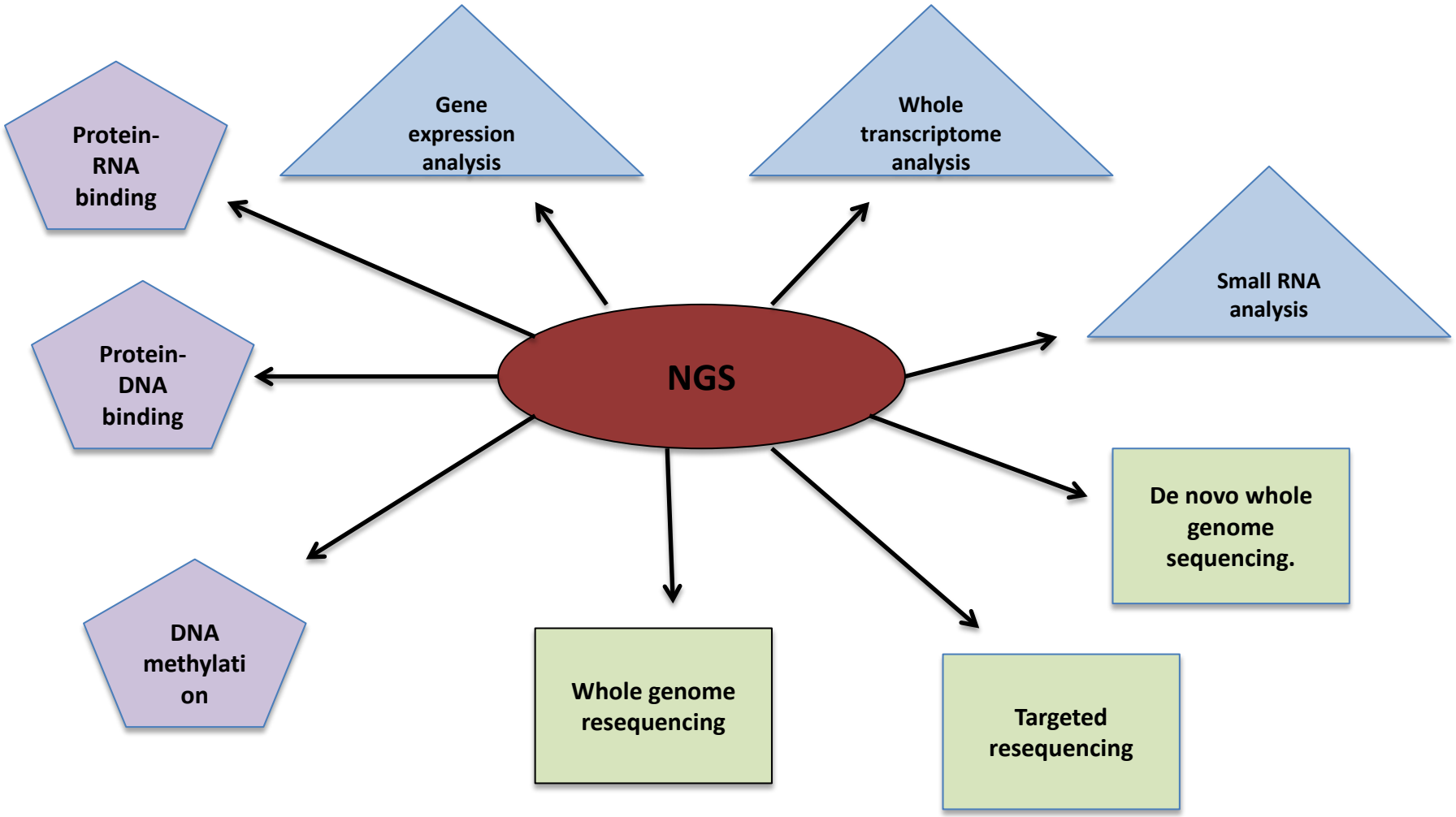


1990s – mid 2005
>800Kbp/day - $1000

~30-40Gbp / lane, 11 days
~$2500/lane
Full flow cell: ~300Gbp
~$18,000

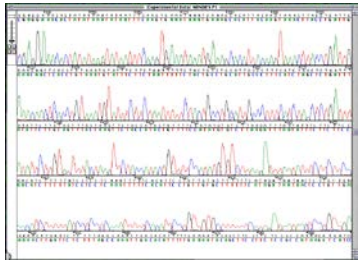
~2Gbp, 8hr
~$1000

Improvements in sequencing has led to an explosion of sequence based experimentation



What does 30Gbp of sequence represent?

Recall:

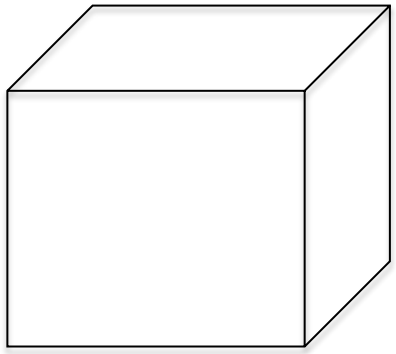


Illumina ~30-40Gbp / lane, 11 days
~$2500/lane.

What if we wanted to print this?

Assuming 2500 characters on a sheet of paper ->
12M pages = 24,000 reams, = 2400 cases of paper!

Dimensions of a case of paper are 12” x 10” x 18”
2400 cases of paper fills a room ~ 14’ X 14’ X 14’

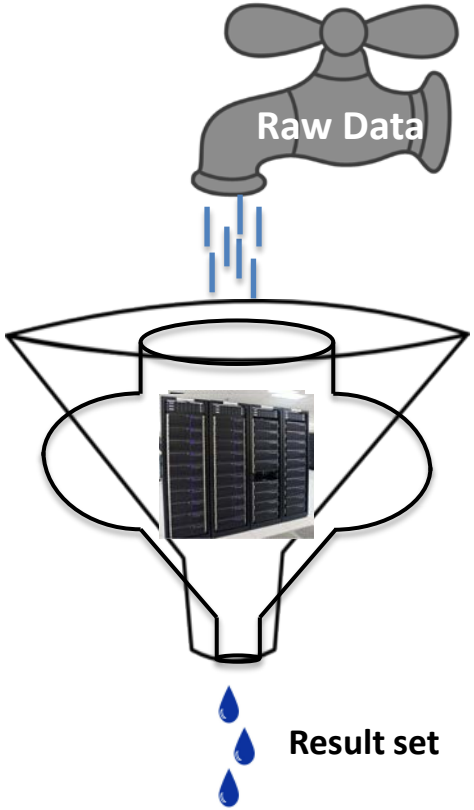


Assuming a fast printer (50pg/min), with a cartridge life of 24,000 pages: 166 days, 500 cartridges -> $246,000

Data Usage Requirements

- High volumes of DNA sequence is accessible to any lab (several contract sequence centers).
- Is a driver of discovery in the life sciences – it use is PERVASIVE, and INCREASING!

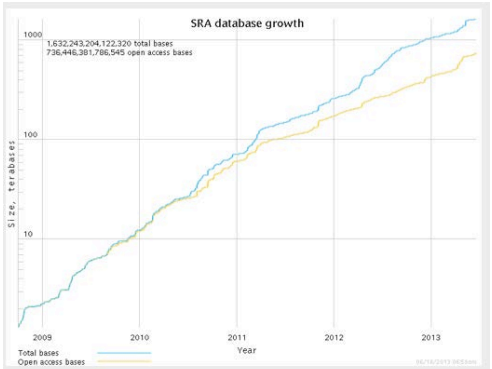
- **RAM and CPU.**
  - NGS is composed of a high volume (100s of millions) of short sequences.
  - Assembling these is very RAM intensive.
  - Manipulating these in a reasonable time-frame is CPU intensive.



- **Storage.**
  - Raw data volume is high.
  - Data footprint increases substantially during processing.

Concluding remarks

- Availability of NGS results in DNA sequence underpinning many lines of inquiry.



— The number of short read datasets available from the NCBI SRA skyrocketed in the last few years.
(Source: http://www.ncbi.nlm.nih.gov/Traces/sra/i/g.png)

- NGS gets cheaper, sequence volume increases, lengths increase.
- Increased lengths improve assembly, but volume will still challenge RAM. Short read applications expanding – challenges CPU and storage capacity.
- NGS moving rapidly into clinical applications, which presents great opportunities for UF.