Data Challenges in the Genomic Research

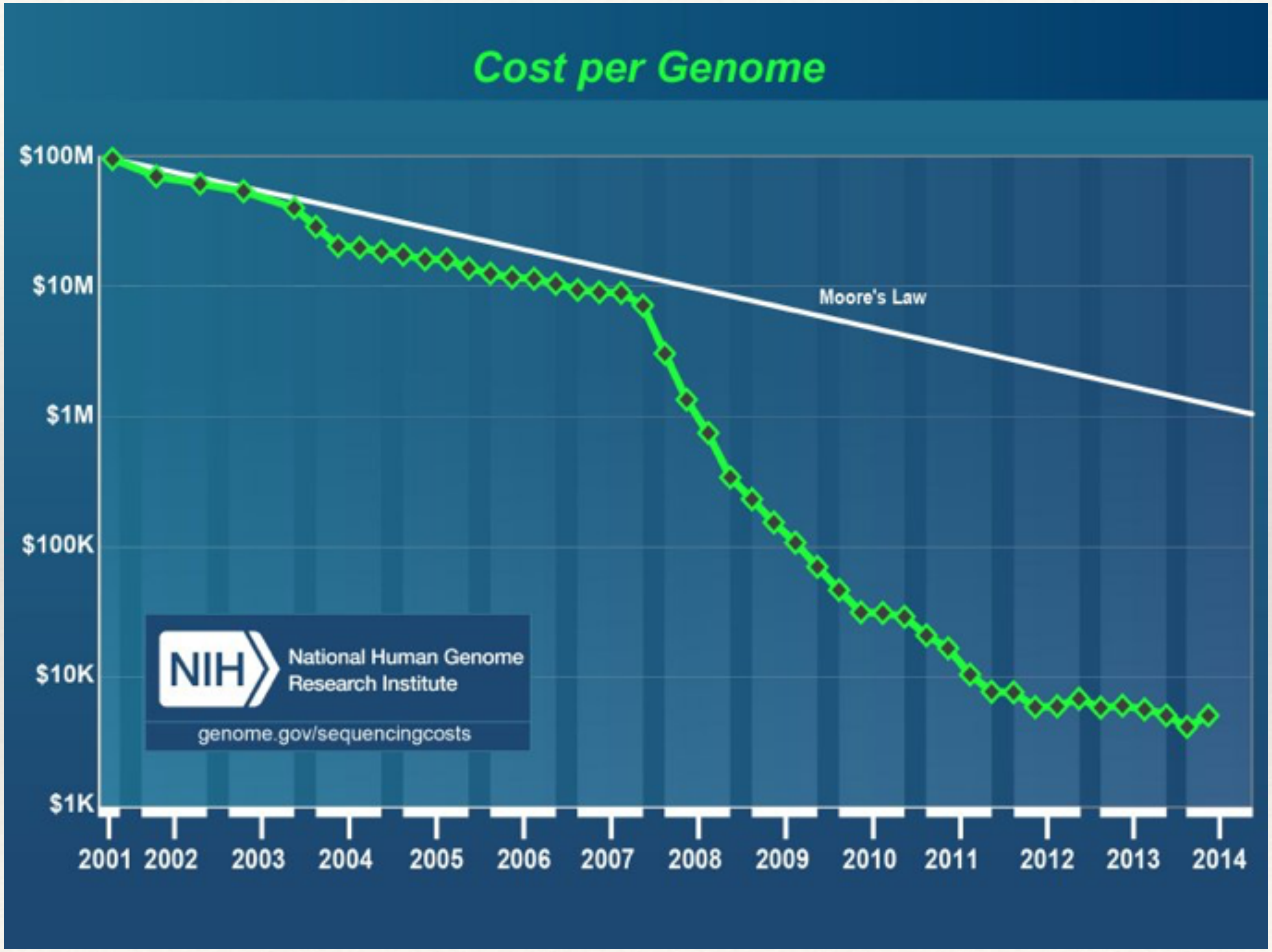
Ian Fisk
Simons Foundation

Community

- The Simons Foundation supports a lot of kinds of science, but they have a large program in Autism research
    - A powerful tool for this is genetic sequencing

What Changed

- Genetic sequencing cost has decreased exponentially



Files

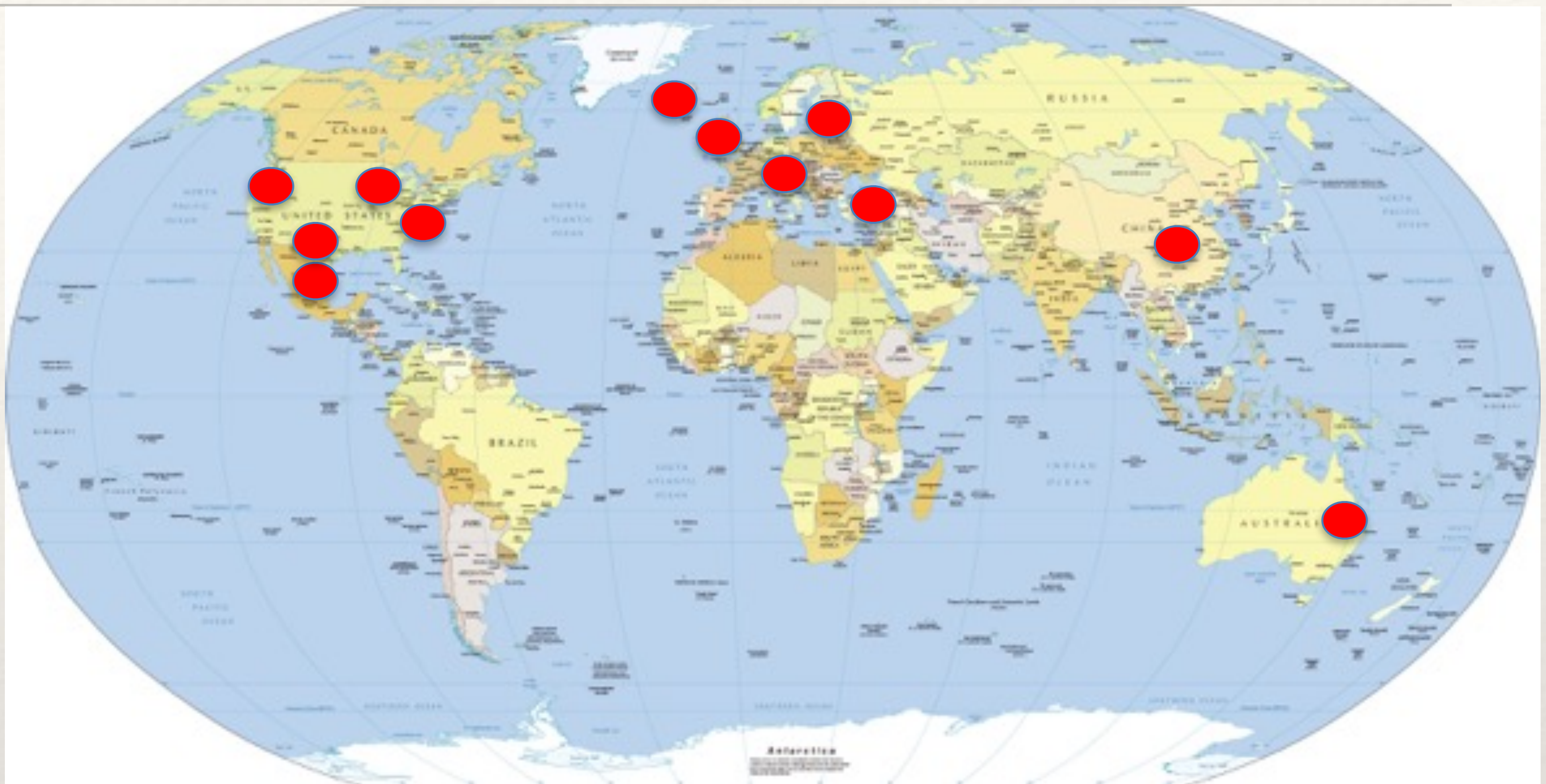
- The machines produce a reasonable standard raw data format called BAM files
    - ~10GB file per person for whole exome (A study of about 1% of your DNA) Mutations here can have severe impact on the rest
    - ~200GB file per person for whole genome sequencing. Modern machines can sequence the entire genome
- A study of a group might be a few hundred individuals
    - Raw data in the few TB range for exome and few hundred TB for full genome to a few PB
- What had been manageable, just became unmanageable

What is different?

- In physics a lot of central effort is used to reduce the data format from raw to summarized analysis formats
  - This is also true of genomics, but this is done much more often by users and groups
- Instead of a 200GB file being a sequence of independent small events, it is sequence of DNA that all has to be analyzed and compared together

Current Distribution

- Currently this genomics community uses FNAL as an archival system
  - Recently imported ~400TB of data primarily from S3
  - 2 100TB samples were exported from FNAL using GridFTP to Iceland and Oregon for additional processing



- The community created and made publicly available 11TB of diversity project data
  - 300 people from all over the planet

The Challenge (1/2)

- There are about 40 entities that want samples of about 1PB
- There is no real infrastructure for data management
  - File lists are sent with checksums in manifests
- These are labs with firewalls and data has grown much faster than expertise, so little community knowledge for how to move big samples around
  - Bare GridFTP is not completely user friendly nor is the entire grid certificate infrastructure

The Challenge (2/2)

- Path through the files in question is a semi-pathological
    - Unpacking internal buffers and retrieving objects across large swaths of the file
        - Access through the file during analysis is not linear and applications know nothing of training or pre-fetching

Data Handling

- One aspect that is clearly lacking is tools for distributing files and providing access to data
  - Datasets are defined by manifest lists (text files)
  - Where data physically is documented on web pages
- Works for a limited amount of data but will not scale for long

How bad is it?

- Recently the sequencing center announced the first 125TB were ready to ship
    - We were getting 50% duty cycle on a 1Gb/s link
        - 4TB/day so more than a month to transfer
- Users got their pitchforks and I got a taxi

Data Consistency

- These files are 200GB each
  - An MD5 Checksum takes hours
  - We discovered recently the files are internally gzipped into buffers and you can check consistency in parallel by checking the validity of the gzips

Collaborating

- Physics Communities have been dealing with high volumes of data and global distribution for years
- The access patterns are different but it would be interesting to look at remote data access