# Issues in Data Storage and Data Management in Large-Scale Next-Gen Sequencing

Matthew Trunnell

Manager, Research Computing

Broad Institute

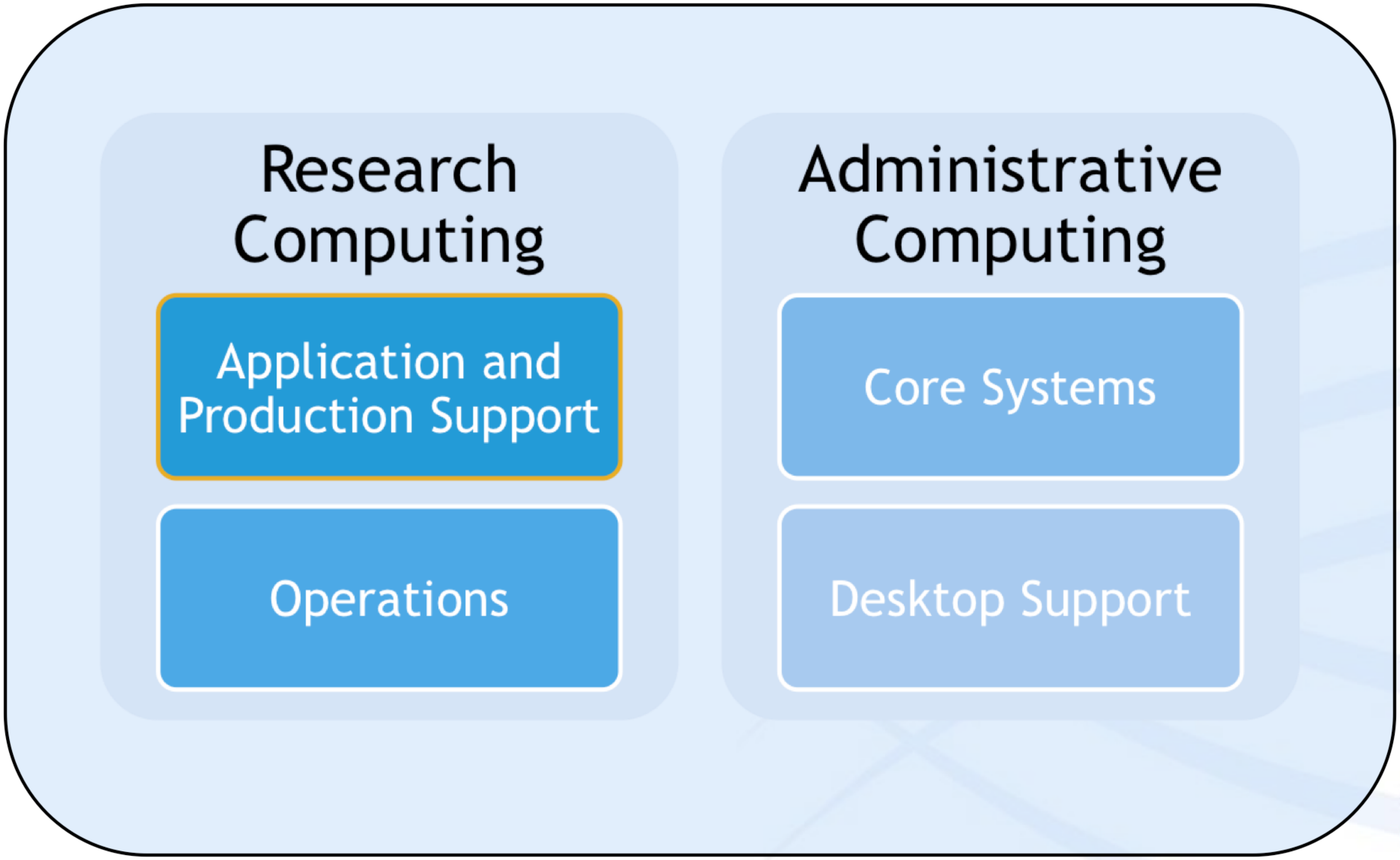
# Overview

- The Broad Institute
- Major challenges
- Current data workflow
- Current IT resource budgeting
- Other operational concerns

# The Broad Institute

- Launched in 2004 as a "new model" of collaborative science with the goal of transforming medicine through genomic research
- An independent, not-for-profit research institute
- Community of ~1400 and growing
- Large-scale generation of scientific data
  - Genomic sequencing
  - Genotyping
  - High-throughput screening
  - RNAi, proteomics
- Scientifically (and computationally) diverse

# Broad IT organization



Research Computing

Application and Production Support

Operations

Administrative Computing

Core Systems

Desktop Support

# Broad Sequencing Platform

- Genome sequencing represents the largest single effort within the Broad, with more than 175 employees overall
- The largest of the Broads' seven data-generation "platforms"
- The Sequencing production informatics team numbers 28, and is responsible for
  - LIMS
  - Production data analysis
  - Production data management
- A service organization operating on a cost-recovery basis
- A major NHGRI Sequencing Center

# Major Challenges

- Data storage
- Data management
- Defining deliverables

# Data Storage

- Next-gen sequencing technologies generate a large amount of data. In the face of this influx of data, how does one:
  - Effectively plan capacity requirements when the sequencing technology and scientific applications are evolving so rapidly?
  - Avoid overprovisioning?
  - Organize data into flexible namespaces to accommodate changing needs?
  - Provide effective protection of data against disaster and accidental deletion?

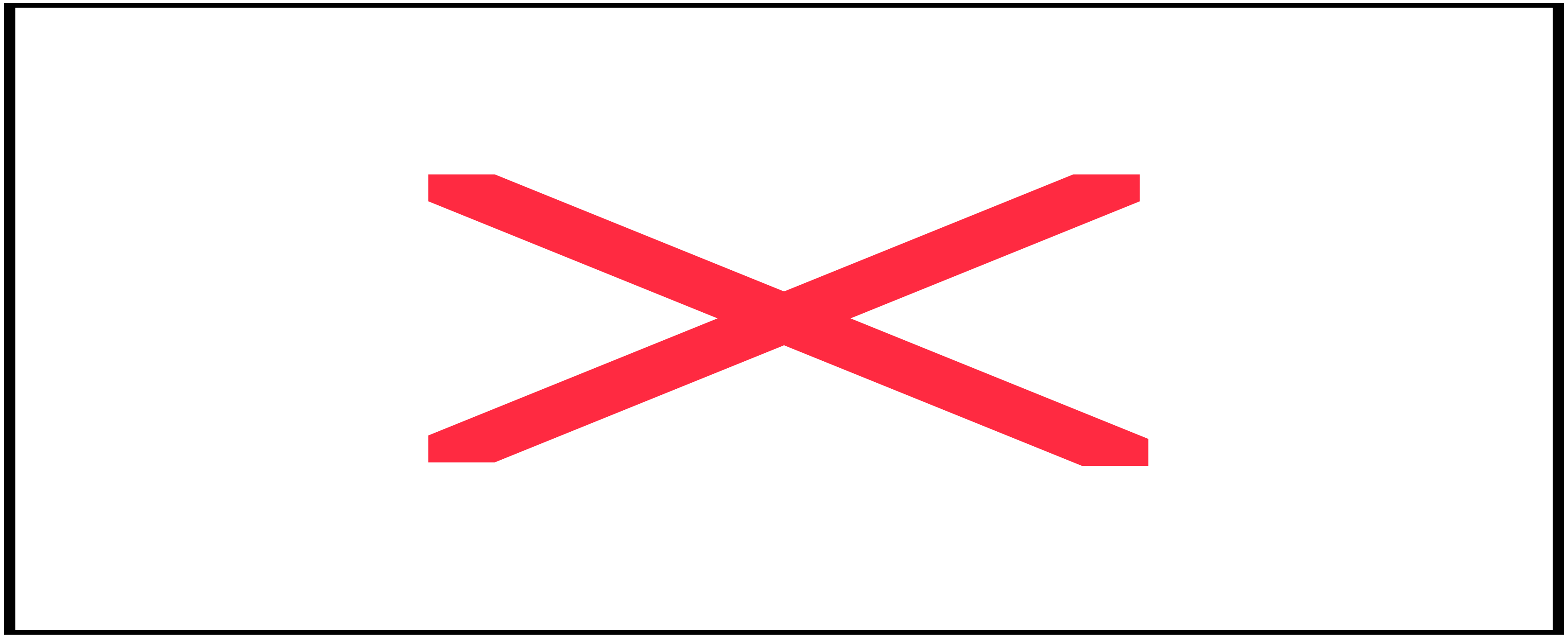
# Data Management

- Data management comes down to defining the life-cycle for the different data files and automating the imposition of that life-cycle:
  - Which of the various data files need to be retained and for how long?
  - How does one confidently automate deletion of intermediate data files?
  - How does one address the needs of special experiments (e.g., rare samples, new protocols)?

# Defining Deliverables

- For many researchers, a FASTQ file with tens of millions of short reads is not immediately useful.
  - How much analysis should be performed “upstream” as part of data generation?
  - To what degree should data be reduced/aggregated?
  - What is the most useful format for delivery of data to researchers?
- While primarily an informatics issue, these questions have direct impact on capacity planning for compute and storage.

# Generalized Data Flow



# Data Storage Philosophy

- Generally we have moved to separate data (into different namespaces, perhaps on different classes of file server) according to its lifetime on disk
- True DR protection has been considered fiscally extravagant for the most part

Raw data

# Raw Image Data

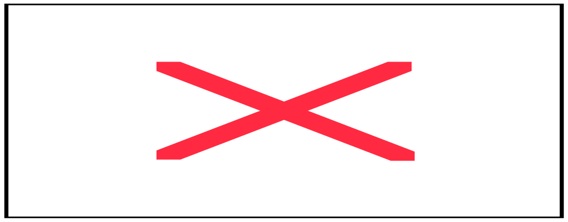
- Discarded immediately after processing, except for special runs:
  - Rare samples
  - New protocols
- Ideally, never leave instrument PC
- Subsampled for process QC
  - Stored as JPEG (or planned to be)

Discarding primary data represents a fundamental shift in how we think about data

Raw data

# Data Storage: Raw Data

- When they are kept, image data are stored on SunFire x4540 “thor” file servers
- File system snapshots provide protection against accidental deletion
- No backups are performed



Intermediate data

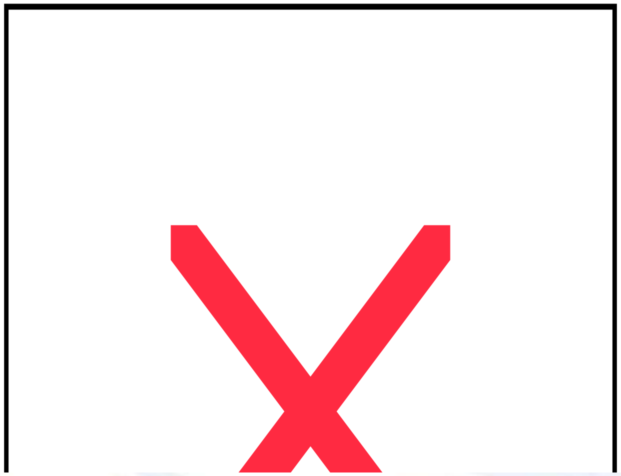
# Intermediate Data

- In-process data represent 1.5 petabytes in our network storage infrastructure, even those these data are budgeted to have a life span of only 30 days.
- .int and .nse are the bulkiest. In theory these can be discarded after base calling. In practice, we use .int files to recalibrate prior to alignment, so they are kept for the full duration of production analysis.

# Data Storage: Intermediate Data

Intermediate data

- In-process data are maintained on two large Isilon clusters, based on X-series node-pairs (24T)
- File systems are not snapshot protected
- Data are not backed up



Processed data

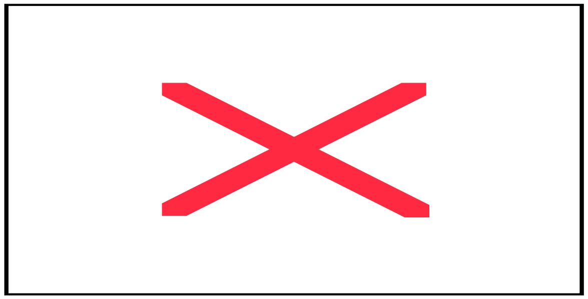
# Processed Data

- The FASTQ files ("sequence.txt") and associated output files from Gerald represent the primary output of the sequencing pipeline
- These data are considered permanent, and are intended to be archived "forever"
- "Forever" == 5 years
- These data are not generally useful to most downstream researchers

Processed data

# Data Storage: Processed Data

- Stored on Isilon NL-class cluster
- Mirrored to low-tier storage (Sun Thumper) for DR purposes



Aligned data

# Aligned Data

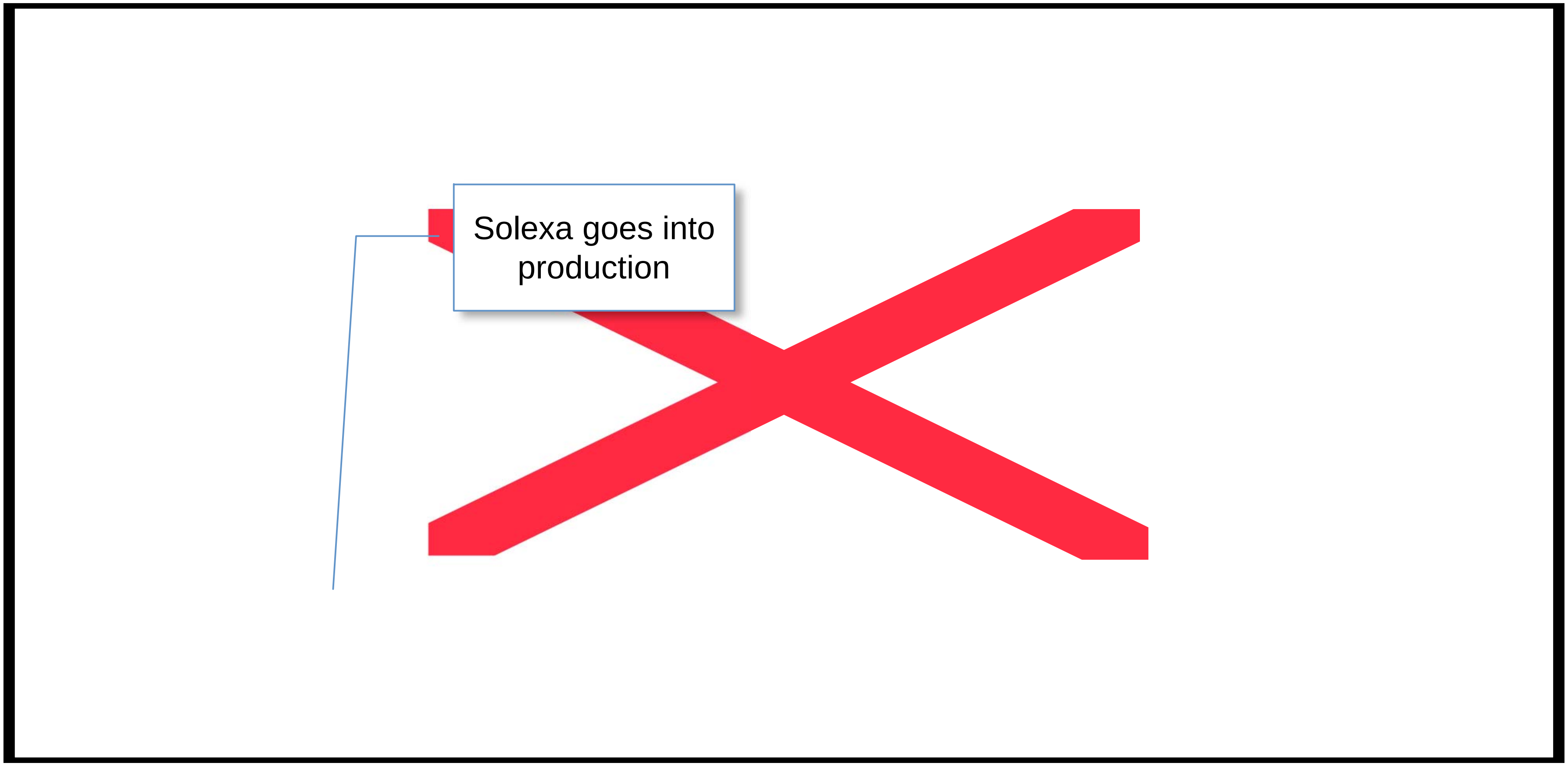
- MAQ is present aligner of choice
- SAM/BAM has become the de facto standard file format among sequencing centers for storing and distributing aligned data
- BAM files containing what?
    - per-lane data, both aligned and unaligned
    - per-library data
    - per-sample data
- Stored online “forever”

Aligned data

# Data Storage: Aligned Data

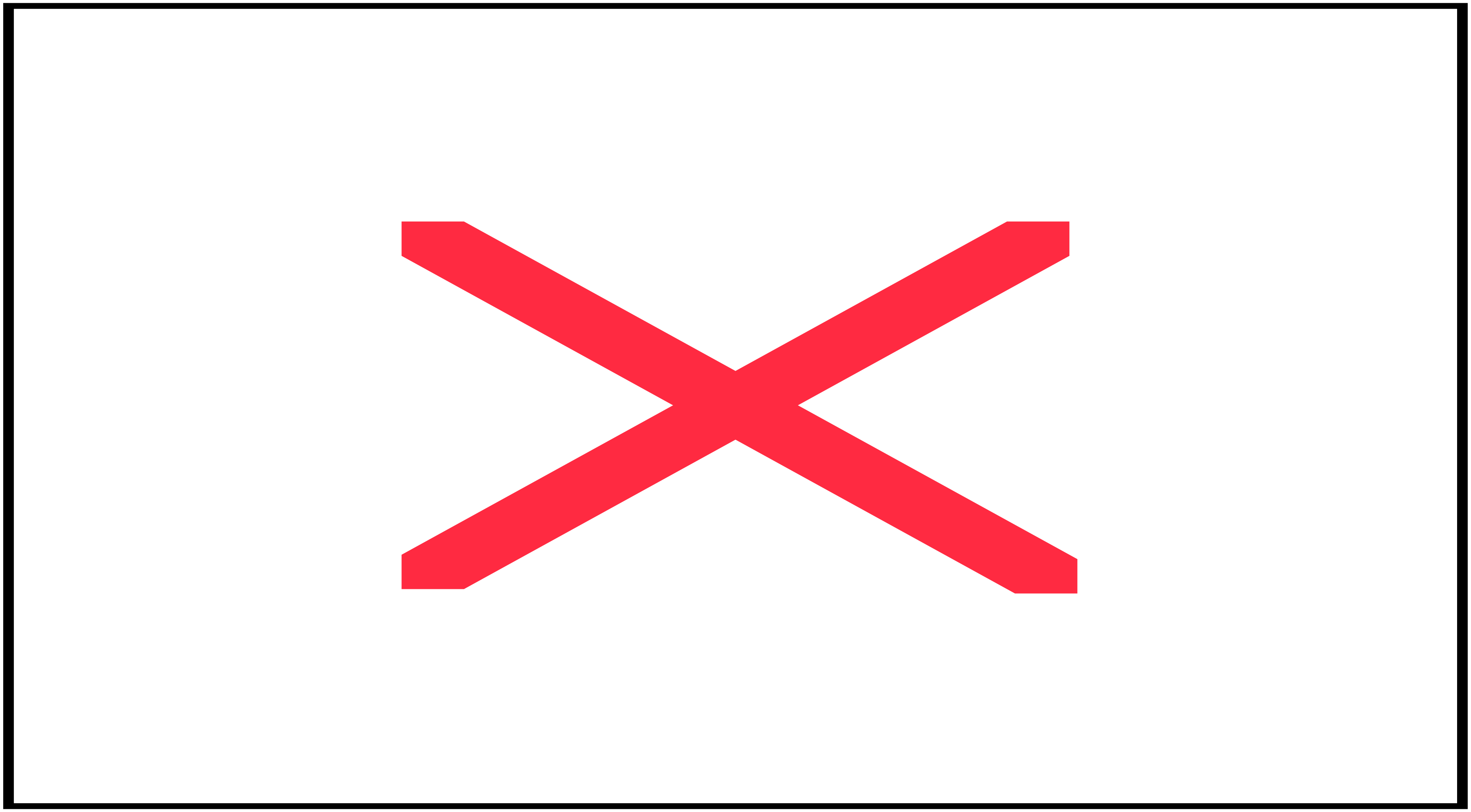
- Stored on Isilon clusters, presently on X-series hardware but plan to scale out on NL-class hardware
- Sequencing informatics just deploying locally-developed content management system to provide access to processed/aligned data (BAM files)
- This may evolve to stand between the end of the analysis pipeline and the final storage pools, allowing more flexibility in managing where those data are stored

# Growth in Sequencing Storage



Solexa goes into production

## Sequencing dominates storage at Broad



# Data Storage

- Why Isilon?
  - Low cost of administration
  - Ease of just-in-time deployment
  - Large namespace
- Why Thumper?
  - Cost low enough to be considered disposable storage
  - But: high cost of administration

# Projecting Storage Requirements

- Think in bases, not bytes
- Think per day, not per run

→ Key planning metric is Gbase/day

# Current IT Resource Budgeting

- For each Illumina GAIIx
  - Two 8-core 32GB blade servers
  - 10-20T of space for intermediate data storage (10T/month retention)
  - Cost of intermediate data storage amortized with cost of instrument
- For long-term storage
  - ~10 bytes/base sequenced (3-4 bytes/base BAM files, ~4 bytes/base genotype data)
  - Cost of long-term data storage passed directly to research grants

# General Suggestions

- Design for flexibility
  - Sequencing technology and sequencing informatics evolving more rapidly than IT
- Define and enforce data life cycles
- Develop good relations with your storage vendor
- For budgeting consider long-term storage as a “consumable”

# Most Common Operational Issues

- Running out of disk space on instrument PC
- Running out of space on network storage for intermediate data
- Transient analysis pipeline failures
- Instrument PC failures

# Monitoring Data Collection

- We have implemented a GlassFish-based application infrastructure with a small client that runs on each data collection PC
- The client:
  - Monitors local disk space
  - Log events from Illumina data collection software with GlassFish server
  - Supports transfer of image files (using bittorrent)
- The server logs state to a central database

# Acknowledgements

- Application Production and Support Group
  - Jean Chang
  - John Hanks
  - Michelle Campo
- Sequencing Informatics
  - Toby Bloom
  - Nathanial Novod
- Sequencing Operations
  - John Stalker