

Answering the demands of digital genomics

Experts Panel

Sept 19, 2011

Beyond the Genome

Genomics beyond 2011

- The cornerstones of genomics continue to be *observation*, *experimentation*, and *interpretation* of the living world
 - Technology has and will continue to push the frontiers of genomics
 - Measurements will be made *digitally* in great quantities, at extremely high resolution, and for diverse applications
- Demands of digital genomics
 1. *Experimental design*: selection, collection, tracking & metadata
 - Ontologies, LIMS, sample databases
 2. *Observation*: measurement, storage, transfer, computation
 - Algorithms to overcome sensor errors & limitations, computing at scale
 3. *Integration*: multiple samples, multiple assays, multiple analyses
 - Reproducible workflows, common formats, resource federation
 4. *Discovery*: visualizing, interpreting, modeling
 - Clustering, data reduction, trend analysis

Sequencing Challenges

- Overcome sequencing limitations through smarter algorithms
 - Co-development of protocol and computational methods
 - Can't sequence entire genomes -> Whole genome shotgun assembly
 - Reads have sequencing errors -> model error types, correct for them
 - Mate-pair protocols fail -> filter redundant pairs, failed mates
- Sequencing frontier:
 - HiSeq 2000: 600 Gbp / run, ~2% error rate, 100bp reads
 - PacBio RS: 150 Mbp / run, ~15% error rate, 1kbp+ reads
- Algorithms frontier:
 - Error correction, deeper coverage
 - Improved indexing, backtracking search, etc

Computing Challenges

- Overcome computing limitations through parallel computing
 - Sensors improving faster than processors, using multiple processors at once
 - GNU Parallel is my new favorite command, limited by cores
 - Batch systems well established for embarrassingly parallel computation, limited by algs.
 - Hadoop, MPI, etc for more flexibility, limited by tools
- Computing frontier:
 - Quad-XL: 8 cores (16 HT), 23 GB RAM, 1.6TB disk => \$1.6/hr
 - Commodity: 4 cores (8 HT), 2.8GHz, 24 GB RAM, 2TB disk => \$2k
 - HighMem: 24 cores (48 HT), 2.0 GHz, 512 GB RAM, 6TB disk => \$35k
 - Blacklight: 4096 cores, 2.27 GHz, 32TB RAM => \$2.5M

Storage and Transfer

- Overcome storage & transfer limitations through improved technology
 - Compress, filter, throw away
 - Transfer: Buy higher capacity internet, use smarter protocols
 - Storage: Buy higher capacity disk, parallel file systems, tiered storage
- Storage frontier:
 - Very large data volumes: Isilon OneFS – 15.5 Pbp in a single volume
 - Very large total capacity: BlueArc – 16 Pbp in a single namespace
 - Parallel distributed filesystem: Lustre – 10 Pbp+, 100+ Gbps
 - Commodity HDFS: 5Pbp+, \$250 / TB
 - Commodity RAID: 24TB / \$9000
- How do we balance convenience of large volumes with technical demands of supporting many concurrent users, replication times, etc

Thank You!

<http://schatzlab.cshl.edu>
[@mike_schatz](#) / [#btgII](#)