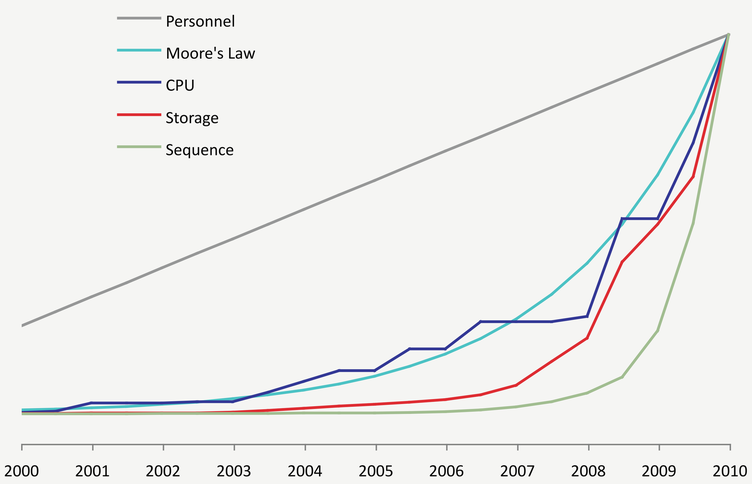
MAXIMIZING UTILITY OF GENOME SEQUENCE DATA

David J. Dooling, Lynn Carmichael, Li Ding, Craig S. Pohl, Scott Smith, Joshua McMichael, George M. Weinstock, Elaine R. Mardis, and Richard K. Wilson
The Genome Center at Washington University in St. Louis School of Medicine

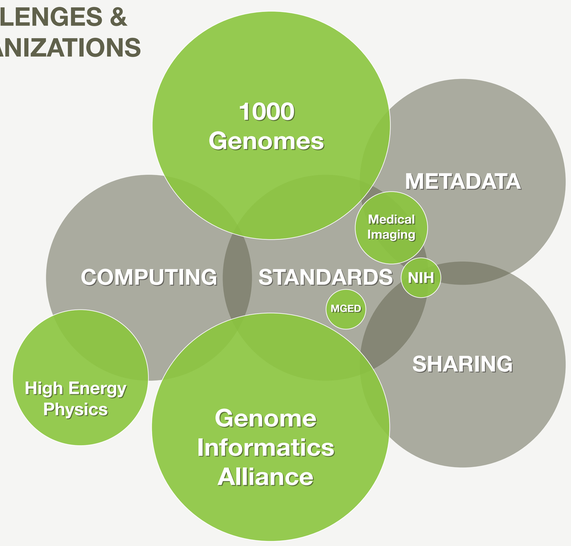
ABSTRACT

Advances in DNA sequencing technologies over the past few years have led to increases in data generation and processing rates that far outpace Moore's Law and storage capacity improvements. As a result, there will come a time when one will no longer be able to "throw more money" at the problems presented by DNA sequencing, i.e., researchers will not be able to keep pace with data generation by purchasing more and more storage and computational nodes. Proposed sequencing platform improvements and the rapid rate of adoption of these technologies by labs large and small will only hasten the time when the old solutions will no longer apply. The history of freely shared sequence data through the NCBI and EBI Trace Archives transform the very difficult problem of massive sequence data generation into a problem of data generation and data sharing on a scale heretofore unimaginable. Over the last year, several organizations, e.g., MGED, NCI, Illumina, 1000 Genomes DCC, and NHGRI, have convened meetings to discuss the problems presented by the massive amounts of data generated by next-generation sequencing technologies. As prologue, brief overviews of these meetings will be presented along with approaches to dealing with massive data generation rates from other disciplines, e.g., high energy physics and high-resolution medical imaging. The Genome Center at Washington University in St. Louis, due to its large-scale sequencing operation and whole-genome analysis capabilities, experiences the difficulties presented by massively-parallel sequencing platforms acutely. To address the many challenges presented by the scale of data generation and requisite analysis, we have developed a multidisciplinary approach involving experts in biology, genomics, bioinformatics, computer science, information technology, and engineering. The resulting approach involves many techniques including intelligent compression and data reduction, data aging, archiving, parallelization, fault-tolerant workflows, scalable software frameworks, and multivariate / multi-genome visualization and comparison, which leverage and extend our laboratory information management system. This approach and its application to the sequencing and analysis of cancer samples will be presented.

MOORE'S LAW



CHALLENGES & ORGANIZATIONS



UR

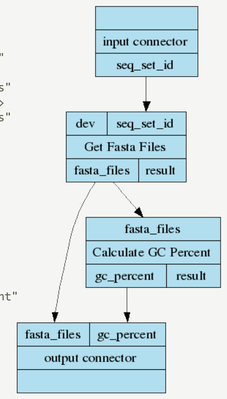
- Object Relational Mapping (ORM) layer
- Distributed, in-memory transactions (STM)
- Manage object state between application and persistence layer (RDBMS)
- Highly parallel and high throughput
- Written in Perl

Genome :: Workflow

```
<?xml version='1.0' standalone='yes'?>
<workflow name="Example Workflow">
  <!-- links -->
  <link fromOperation="input connector" fromProperty="seq_set_id"
   toOperation="Get Fasta Files" toProperty="seq_set_id" />
  <link fromOperation="Get Fasta Files" fromProperty="fasta_files"
   toOperation="Calculate GC Percent" toProperty="fasta_files" />
  <link fromOperation="Get Fasta Files" fromProperty="fasta_files"
   toOperation="output connector" toProperty="fasta_files" />
  <link fromOperation="Calculate GC Percent"
fromProperty="gc_percent"
   toOperation="output connector" toProperty="gc_percent" />

  <!-- operations -->
  <operation name="Get Fasta Files">
    <operationtype commandClass="MGAP::Command::GetFastaFiles"
     typeClass="Workflow::OperationType::Command" />
  </operation>
  <operation name="Calculate GC Percent">
    <operationtype commandClass="MGAP::Command::CalculateGcPercent"
     typeClass="Workflow::OperationType::Command" />
  </operation>

  <operationtype typeClass="Workflow::OperationType::Model">
    <inputproperty>seq_set_id</inputproperty>
    <outputproperty>fasta_files</outputproperty>
    <outputproperty>gc_percent</outputproperty>
  </operationtype>
</workflow>
```



Genome :: Model

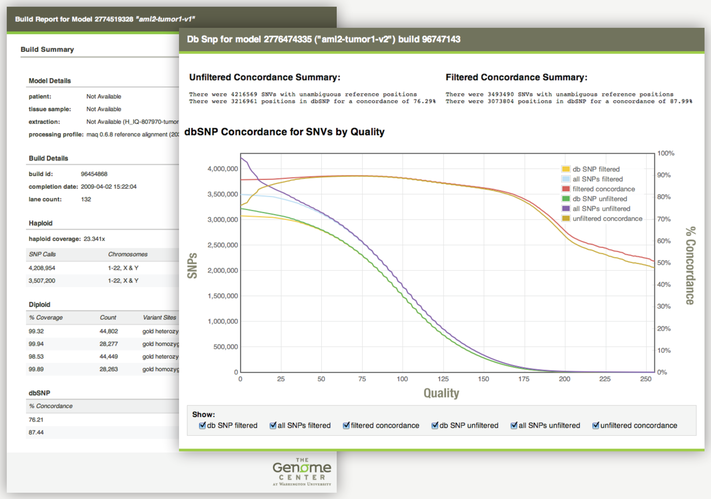
- Represents a state of belief about a genome
- Genome model information includes:
  - Patient/sample phenotype
  - Sequence, array, etc. data
  - Software versions and configuration
  - Analysis results
- A single sample/patient/population could have several concurrent models
- Built on UR and Genome::Workflow

Genome :: Model :: Tools

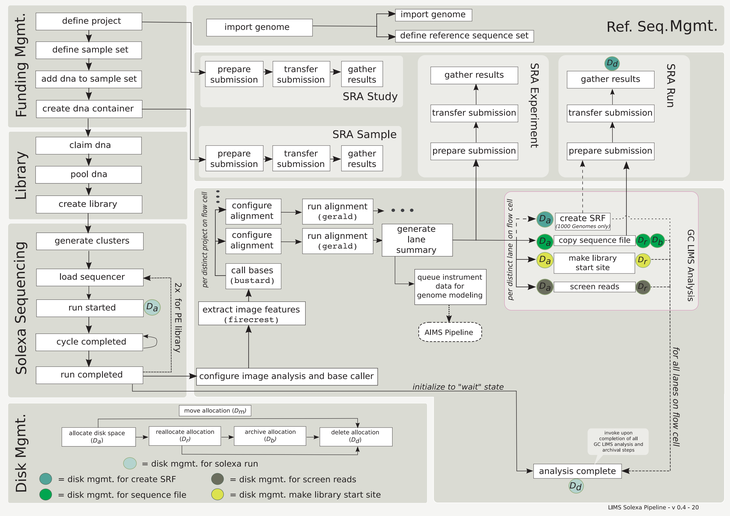
```
1 #!/gsc/bin/perl$
2 $
3 use strict;$
4 use warnings;$
5 $
6 unless( $ARGV[0]) {$
7     print "Usage: wc.pl filename\n";$
8     exit 1;$
9 }$
10 $
11 unless ( -e $ARGV[0]) {$
12     print $ARGV[0] . "Doesn't exist!\n";$
13     exit 1;$
14 }$
15 $
16 my $cmd ='wc -l';$
17 my $return = `$cmd  $ARGV[0]`;$
18 print $return , "\n";$
19 exit;$
```

```
1 package Genome::Model::Tools::WordCount;$
2 $
3 use strict;$
4 use warnings;$
5 $
6 use Genome;$
7 $
8 class Genome::Model::Tools::WordCount {$
9     is => 'Command',$
10    has => [$
11        filename => {is => 'String', doc =>'the file to be wordcounted'},$
12    ],$
13 };$
14 $
15 sub execute {$
16     my $self=shift;$
17     my $filename=$self->filename;$
18     unless ( -e $filename) {$
19         print $filename . "Doesn't exist!\n";$
20         return;$
21     }$
22     my $cmd ='wc -l';$
23     my $return = `$cmd  $filename`;$
24     print $return , "\n";$
25 }$
```

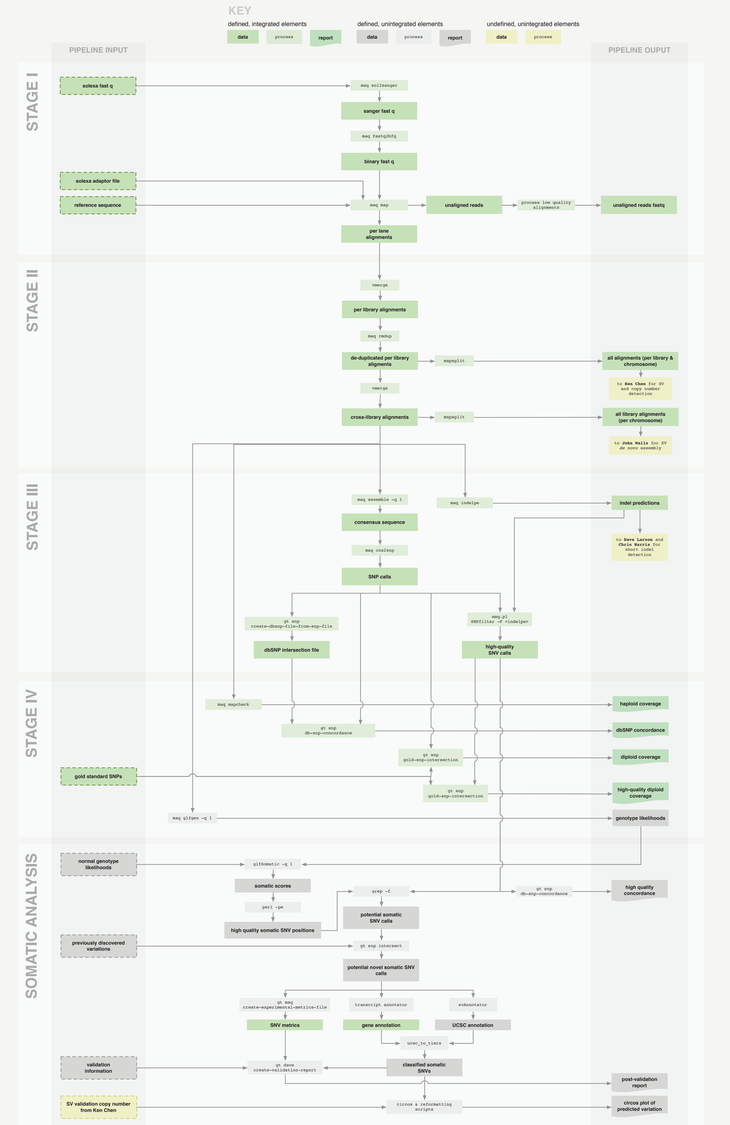
Genome :: Model :: Reports



LIMS SOLEXA PIPELINE



VARIATION DETECTON PIPELINE



FUTURE WORK

- Public release (GPL3)
- Improved reports
  - Pipeline status and result reports
  - Enhanced genome-model comparison tools
- High-throughput computing

Open Science Grid

BOINC

amazon web services

THE Genome CENTER AT WASHINGTON UNIVERSITY