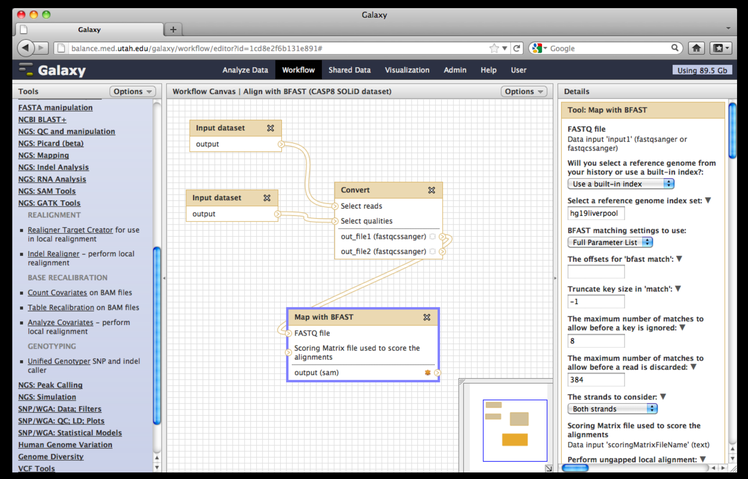
Making Sense of the Next-Gen Genetic Sequencing Pipeline *with Galaxy*

Bowtie
MAQ
Alignment
BWA
Novoalign
BFAST
BioScope
Picard
MarkDuplicates
AddOrReplaceReadGroups
Post-processing
samtools
RealignerTargetCreator
IndelRealigner
AnalyzeCovariates
CountCovariates
GATK
TableRecalibration
Variant Calling,
Genotyping
UnifiedGenotyper
VariantFiltration
VariantAnnotator
Annotation
ANNOVAR
PolyPhen
SIFT
Analysis
Savant
IGB
IGV
MedSavant
UCSC Genome Browser
VAAST
Excel Spreadsheets
Visualization?



The Blessings - And Curses - Of NGS Data

While Next-Generation Sequencing (NGS) offers biologists a way to obtain genetic data at a fraction of the cost and unprecedented speeds compared to conventional sequencing, it also comes with unprecedented complexity - imagine trying to re-assemble an encyclopedia from its shredded pages. That's roughly the task that alignment programs are faced with, and it's only the first step in NGS analysis. Hundreds of algorthims have been developed to work with NGS data (a *small* sample is portrayed on the far left), but no tool has yet emerged as a "one size fits all" solution. Each has its own specific function, as well as its own file format, documentation, and parameters - and just about all of them still only have a command-line interface.

Obviously, there is huge potential for human error - *meticulous* data provenance here is key - not to mention the sheer usability issues that a biologist (especially one with little command-line experience) is likely to encounter.

Cleaning It Up, Reducing Human Error

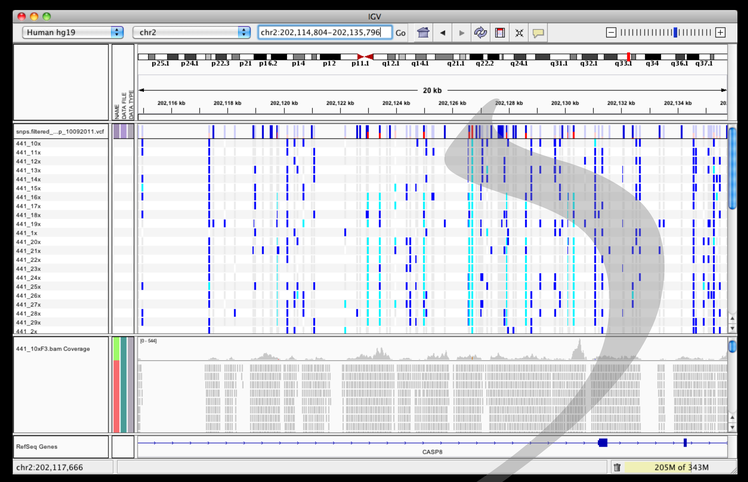
Galaxy[1] is a tool that is being developed at Penn State and Emory Universities, with a very broad, active developer community. It's a web-based tool that provides a GUI front-end to a small handful of these NGS tools. Tools can be chained together graphically, program parameters can be set all on the same screen, there are ways of providing on-screen documentation, and there's a built-in data provenance system that ensures repeatability, as well as making the specifics of a pipeline much easier to publish.

We are currently working on adding more interfaces between Galaxy and existing NGS tools such as Novoalign[2], currently unimplemented GATK[3] modules, ANNOVAR[4], and VAAST[5] (an advanced analysis program that was recently developed by Mark Yandell at the University of Utah).

Once We Have Our Results, What Do We Do With Them?

One other major challenge with NGS data is the sheer volume of information - while representing huge amounts of data in a meaningful way is not a challenge unique to genetics, it certainly has its own twists, often dependent upon the kind of questions a scientist may be asking. For this reason, we feel that a single, monolithic visualization is unlikely to be effective in all scenarios.

We are currently working with Nicola Camp at the University of Utah's Division of Genetic Epidemiology to develop a visualization technique for an ongoing research project searching for for deleterious and protective genetic variations for breast cancer risk in the chromosomal regions surrounding 3 apoptosis genes: CASP8, DR4 and DR5. With diseases as complex as cancer, it's most likely that we will find several small variations (possibly separated by large distances) that work together to increase or decrease risk.



Currently, there are two major ways of looking at NGS data. One techinique is known as a "genome browser" - just like every other step in the NGS pipeline, there are quite a few of these. One of the more popular is IGV[6], displayed on the right. Some of these, such as the UCSC Genome Browser[7], are already integrated into Galaxy. The other common technique is through a table-based layout and filters, such as an Excel spreadsheet or MedSavant[8]. For out current study, neither approach is likely to be effective; there are far too many variants for a genome browser to be much help, and the biases inherent in arbitrary filtering are likely to obscure at least some of the subtle variants that we are looking for. We are currently building unique, specialized visualizations to solve these problems.

1 http://main.g2.bx.psu.edu
2 www.novocraft.com
3 http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
4 http://www.openbioinformatics.org/annovar
5 http://www.yandell-lab.org/software/vaast.html
6 http://www.broadinstitute.org/igv/home
7 http://genome.ucsc.edu
8 http://genomesavant.com/medsavant/download.php