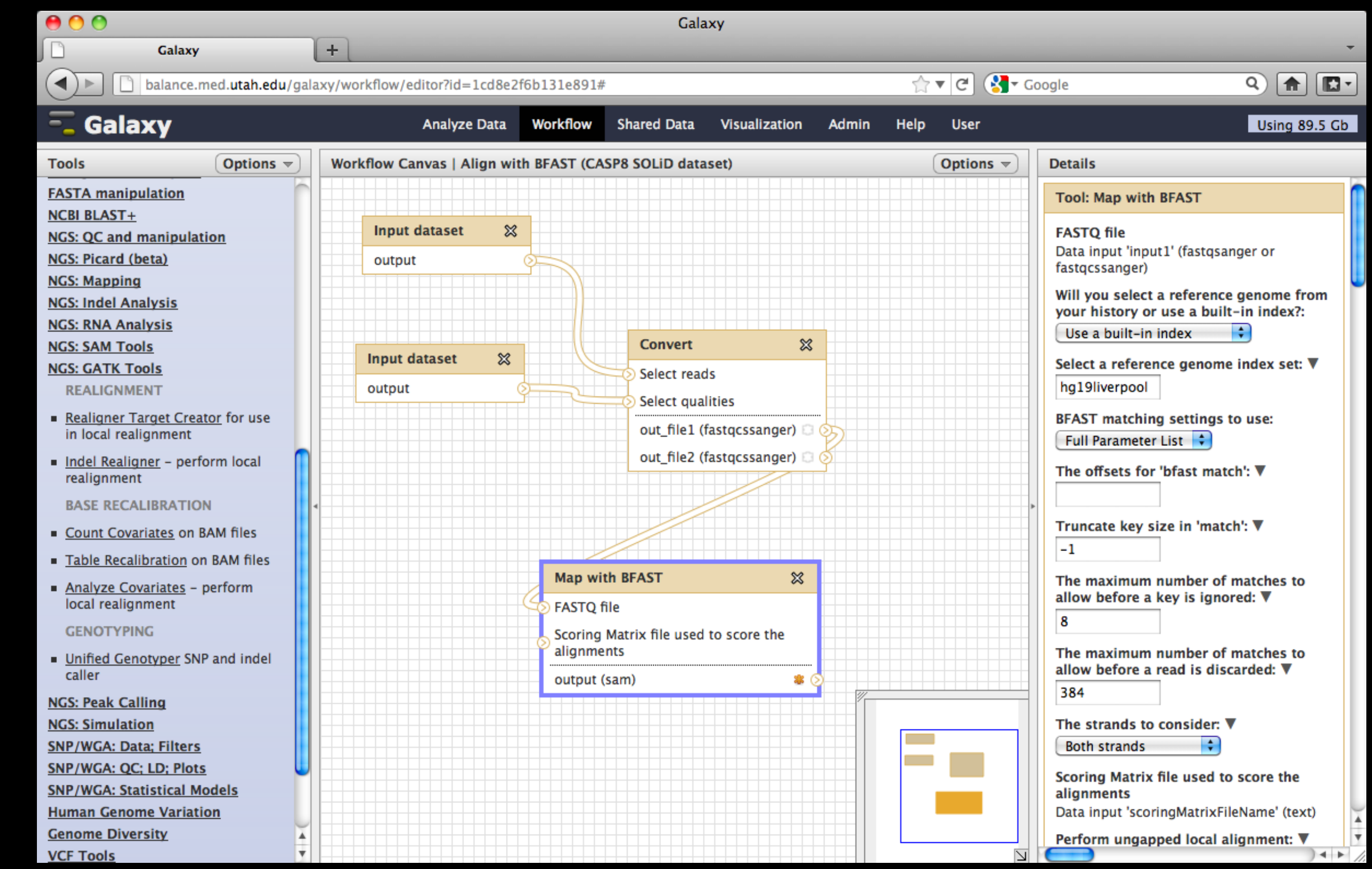


```
usage: java -jar GenomeAnalysisTK.jar -T <analysis_type> [-SM <sample_metadata>] [-rbs <read_buffer_size>] [-et <phone_home>] [-rf <read_filter>] [-R <reference_sequence>] [-B <rodBind>] [-BTI <rodToIntervalTrackName>] [-BTMR <BTI_merge_rule>] [-ndrs <nonDeterministicRandomSeed>] [-D <DBSNV <DBSNV_type>] [-dt <downsampling_type>] [-dfrac <downsample_to_fraction>] [-dcov <downsample_to_coverage>] [-baq <baq>] [-baqGAP <baqGapOpenPenalty>] [-PF <performanceLog>] [-OQ <useOriginalQualities>] [-S <validation_strictness>] [-nt <num_threads>] [-im <interval_merging>] [-rgbl <read_group_black_list>] [-disable_experimental_low_memory_sharding] [-l <logging_level>] [-log <log_to_file>] [-h] [-help] [-glm <genotype_likelihooods_model>] [-pnrmm <p_nonref_model>] [-heterozygosity <heterozygosity>] [-pcr_error <pcr_error_rate>] [-gt_mode <genotyping_mode>] [-out_mode <output_mode>] [-stand_call_conf <standard_min_confidence_threshold_for_calling>] [-stand_emit_conf <standard_min_confidence_threshold_for_emitting>] [-nsl] [-mbq <min_base_quality_score>] [-mmq <min_mapping_quality_score>] [-minIndelCnt <min_indel_count_for_genotyping>] [-indelHeterozygosity <indel_heterozygosity>] [-o <out>] [-debug_file <debug_file>] [-metrics <metrics_file>] [-A <annotation>] [-g <groups>]
```

```
T, --analysis_type <analysis_type>
  SAM or BAM file(s)
  -I, --input_file <input_file>
  -SM, --sample_metadata <sample_metadata>
  -rbs, --read_buffer_size <read_buffer_size>
  -et, --phone_home <phone_home>
  -rf, --read_filter <read_filter>
  -l, --intervals <intervals>
  -XL, --excludeIntervals <excludeIntervals>
  -R, --reference_sequence <reference_sequence>
  -B, --rodBind <rodBind>
  -BTI, --rodToIntervalTrackName <rodToIntervalTrackName>
  -BTMR, --BTI_merge_rule <BTI_merge_rule>
  -ndrs, --nonDeterministicRandomSeed
  -D, --DBSNV <DBSNV_type>
  -dt, --downsampling_type <downsampling_type>
  -dfrac, --downsample_to_fraction <downsample_to_fraction>
  -dcov, --downsample_to_coverage <downsample_to_coverage>
  -baq, --baq <baq>
  -baqGAP, --baqGapOpenPenalty <baqGapOpenPenalty>
  -PF, --performanceLog <performanceLog>
  -OQ, --useOriginalQualities
  -DBO, --defaultBaseQualities <defaultBaseQualities>
  -S, --validation_strictness <validation_strictness>
  -nt, --num_threads <num_threads>
  -im, --interval_merging <interval_merging>
  -rgbl, --read_group_black_list <read_group_black_list>
  -disable_experimental_low_memory_sharding
  -l, --logging_level <logging_level>
  -log, --log_to_file <log_to_file>
  -h, --help

Arguments for UnifiedGenotyper:
  -glm, --genotype_likelihooods_model <genotype_likelihooods_model>
  -pnrmm, --p_nonref_model <p_nonref_model>
  -hets, --heterozygosity <heterozygosity>
  -pcr_error, --pcr_error_rate <pcr_error_rate>
  -gt_mode, --genotyping_mode <genotyping_mode>
  -out_mode, --output_mode <output_mode>
  -stand_call_conf, --standard_min_confidence_threshold_for_calling
  -stand_emit_conf, --standard_min_confidence_threshold_for_emitting
  -nsl, --noSLOP
  -mbq, --min_base_quality_score <min_base_quality_score>
  -mmq, --min_mapping_quality_score <min_mapping_quality_score>
  -deletions, --max_deletion_fraction <max_deletion_fraction>
  -minIndelCnt, --min_indel_count_for_genotyping <min_indel_count_for_genotyping>
  -indelHeterozygosity, --indel_heterozygosity <indel_heterozygosity>
  -o, --out <out>
  -debug_file, --debug_file <debug_file>
```

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



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 algorithms 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.

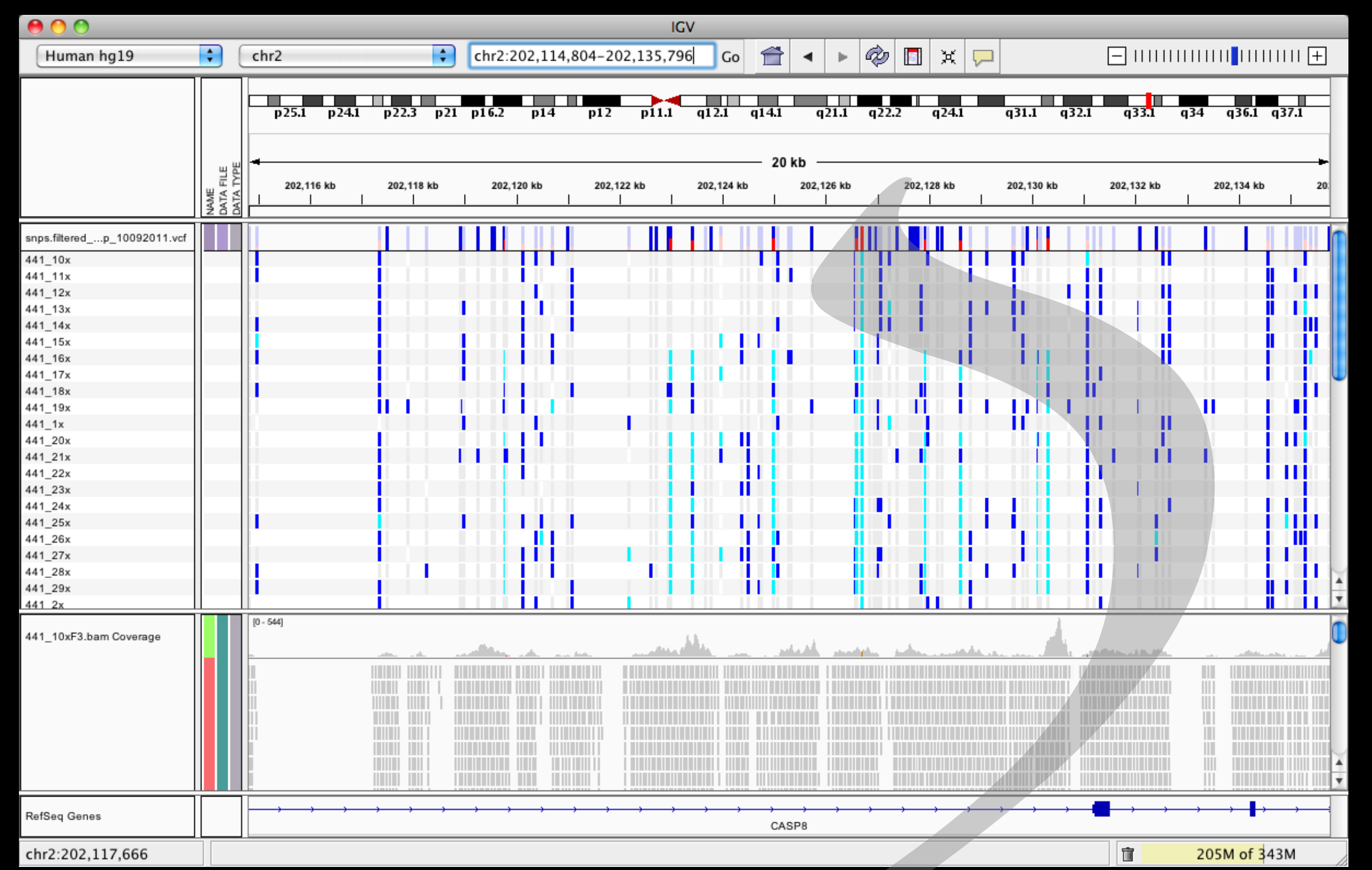
Cleaning It Up, Reducing Human Error

Galaxy¹ 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 Novoalign², currently unimplemented GATK³ modules, ANNOVAR⁴, and VAAST⁵ (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 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 technique 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 IGV⁶, displayed on the right. Some of these, such as the UCSC Genome Browser⁷, are already integrated into Galaxy. The other common technique is through a table-based layout and filters, such as an Excel spreadsheet or MedSavant⁸. For our 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.



```
Genotype likelihoods
  calculation model to employ
  SLOP is the default option,
  while INDEL is available
  for calling indels and BOTH
  available for calling both
  together as INDEL (BOTH)
  Non-reference probability
  calculation model to employ --
  EXACT is the default option,
  while GRID_SEARCH is also
  available. (EXACT|GRID_SEARCH)
  How many threads should be
  allocated to run this tool?
  (DISCOVERY)
  GENOTYPE_GIVEN_ALLELES
  Would we output confident
  genotypes (i.e. including ref
  calls) or just the variants?
  (HIT_VARIANTS_ONLY)
  EMIT_ALL_CONFIDENT_SITES
  EMIT_ALL_SITES
  The minimum phred-scaled
  confidence threshold at which
  variants not at "trigger"
  track sites should be called.
  The minimum phred-scaled
  confidence threshold at which
  variants not at "trigger"
  track sites should be emitted
  (and filtered if less than the
  calling threshold)
  If provided, we will not
  be using SLOP
  Minimum base quality required
  to consider a base for calling
  Minimum read mapping quality
  required to consider a read
  for calling
  Maximum fraction of reads with
  deletions to allow for calling
  (disable, set to < 0 or > 1;
  default:0.05)
  Minimum number of consensus
  indels required to trigger
  genotyping run
  Heterozygosity for indel
  calling
  File to which variants should
  be written
  File to print all of the
```

Savant
IGB
IGV
UCSC Genome Browser
Excel Spreadsheets

Visualization?

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>