Bioinformatics Services

The Bioinformatics group offers support to researchers within UT and outside to assist with management and analysis of large scale data. We use "best practices" highly cited open source tools as well as tools developed within our group for our data analysis.

By having one of our Consultants perform data analysis, you can be assured that:

- Data quality will be interpreted by experienced bioinformaticians
- Any errors in the pipeline will be addressed by experts
- Parameters will be adjusted appropriately (and with your input if needed)
- Additional training and/or interpretation can be provided in the context of your project (at additional cost)
- Pipelines may be customized or extended for a particular project (at additional cost)

Some of the typical services we are capable of offering are listed below. But, please contact us if you do not see something that resembles your particular project. We offer estimates of time and cost wherever it is appropriate, but times and cost can vary based upon level of detail required and specifics of your project. All costs are based on currently approved Service Center rates and the actual amount of labor required for the project. Projects are billed when complete or monthly, whichever comes first. A minimum of four hours of time is assumed for each project. Larger datasets, more complicated experimental designs, or additional interpretation and/or training time can be easily accommodated.

General services available

1. **Benchmarking of tools/pipelines**: New bioinformatics tools are introduced everyday and a thorough comparison is required to select the most appropriate tool. Evaluation of bioinformatics tools for accuracy and performance will be performed using simulated and/or real data.

2. **ddRAD variant calling**: Identification and annotation of variants compared to a reference genome. For higher eukaryotes uses tools such as GATK (Genome Analysis Toolkit) for SNP calling or MuTect2 for somatic mutations. Uses tools such as samtools mpileup or bwa for smaller genomes.

3. **RNA-Seq analysis pipeline**: This pipeline uses an annotated genome/transcriptome to identify differential expressed genes/transcripts. 15 hour minimum ($1095 internal, $1395 external) per project.

4. **RNA-Seq for non-model organisms**: For non-model organisms, a representative transcriptome will be assembled using tools like Trinity. This assembly will be evaluated for completeness and annotated using homology search tools. By mapping to this annotated transcriptome, differentially expressed genes/transcripts will be identified.

5. **Network/Co-expression analysis**: Tools like WGCNA (weighted co-expression network analysis) will be used to identify patterns of correlation in gene/transcript expression in order to identify co-expressed/potentially co-regulated genes.

6. **ChIP-Seq peak calling pipeline**: This pipeline identifies regions of significant protein binding ("peaks") based on an annotated genome. 12 hour minimum ($876 internal, $1116 external) per project.

7. **ChIP-Seq downstream analysis**: Given a set of confident peak calls, this analysis may use a variety of tools to assess biological relevance. Examples include motif analysis, identification of possible regulatory targets, construction of regulatory networks, and differential binding analysis.

8. **Transcriptome assembly**: Assembly of RNA-seq short reads into a transcriptome. 15 hour minimum ($1095 internal, $1395 external) per project.

9. **Genome assembly**: Assembly of genomes using short read data will be performed using tools like Velvet, Allpaths (for specific library types) and SPAdes (for bacterial genomes). If long read technology like PacBio reads are available, they can also be incorporated to improve the genome assembly. The assemblies will be evaluated for completeness and will be annotated using tools like MAKER.

10. **Data visualization**: We can provide custom data visualizations to expand on any of the analyses we provide using a wide variety of R (ggplot, etc.) and python (and other) data visualization libraries.

11. **Promoter analysis**: We have developed in-house an algorithm, SArKS, for de novo discovery of biological sequence motifs and multi-motif domains preferentially present in promoter regions for genes with elevated differential expression scores. We are also happy to apply other standard motif discovery tools (e.g., DREME, HOMER, MEME, MEME-ChIP, etc.) or to help develop custom promoter analysis approaches.

12. **Statistical analysis/biostatistics/machine learning problems**: Downstream analysis of bioinformatics data often requires a number of modern statistical and machine learning methods. We can provide assistance in building, analyzing, and evaluating predictive models (classification, regression, time-to-event, etc.) appropriate for diverse experimental designs (and are also happy to provide advice on experimental design if you are currently planning your next project). We also suggest considering unsupervised multivariate analysis methods (PCA, (N)MDS, t-SNE, clustering, etc.) particularly useful complements to many of the services we provide.

13. **Microbiome community analysis**: Analysis of 16S rRNA sequencing projects (alpha and beta diversity analyses, phylogenetic tree construction, ordination plots, etc.) can be provided using QIIME in conjunction with custom vegan- and phyloseq-based R scripts (along with other tools as desired/necessary for your project). 12 hour minimum ($876 internal, $1116 external) per project.

14. **Application development/Optimization of pipelines to run on HPC environments**: Compute clusters such as those available at the Texas Advanced Computing Center (TACC) offer massive resources for compute-intensive tasks on large data. However, optimizing pipelines to take advantage of the parallel architecture of a compute cluster often requires extra processing steps. We have experience with adapting existing pipelines and software to a massively parallel environment (such as the Trinity transcriptome assembler and BLAST) and can work with researchers to develop pipelines for HPC clusters.

15. **ddRAD analysis**: Double digest RADseq offers a low-cost method for identifying polymorphisms in both model and non-model organisms. We offer ddRAD consulting services centered around the Stacks pipeline, with parameter optimization for the pipeline and parallelization on TACC compute clusters.

Rates
Internal customers (payment from a UT Austin account): $73/hour

External customers (anyone paying from a non-UT Austin account): $93/hour

**Where to go from here**

Email the Consultants at bcg@utgsaf.org with a brief description of your project and analysis needs.