Title: Data Scientist
Reporting: CEO with an indirect reporting relationship to Senior Data Scientist

Candidate: You are passionate about the company's mission to commercialize industrial genomics. You are a self-starter with an inextinguishable fire to compete and succeed. You thrive in an environment that requires crisp judgment, pragmatic decision-making, rapid course-corrections, and comfort with market ambiguity. You enjoy working with complex and diverse data to provide meaningful insights to our customers. You discharge your duties within a culture of mutual team respect, high performance, and humility.

Objectives of this role

- Analyze DNA / RNA sequencing data: assessing quality, cleansing, structuring for downstream processing, analysis and visualization.
- Improve and develop computational pipelines for subsurface and environmental sample DNA data analysis for performance and scalability.
- Collaborate with other technical disciplines and customers to develop new analysis capabilities from proof of concept to commercial use.
- Generate actionable insights for our customers.

Responsibilities:

- Design and execute, computational analysis of environmental DNA data to enhance decision-making of current and future customers of industrial genomics in the energy transition, Oil & Gas, (waste)water, and asset integrity industries. Analysis will require both amplicon-based sequencing and shotgun metagenomic methods and utilize existing Biota pipelines and the development of new methods.
- Develop Python-based computational pipelines for subsurface DNA data analysis, incorporating open-sourced tools (e.g. QIIME, SourceTracker, Emperor, Scikit-learn, Scikit-bio, PICRUSt, etc.) to increase performance and scalability of analysis.
- Participate in execution of projects with geologists and petroleum engineers to provide commercial value for oil and gas customers by integrating genomics derived insights with Oil & Gas data sets.
- Participate in design and implementation of state-of-the-art bioinformatics pipeline to process all forms of DNA sequence data (target genes, shotgun metagenomics, microarray) to enhance analysis and service capabilities.
- Facilitate the development of a pipeline for genome wide variant tracking including organization of the database architecture.
- Develop procedures for data storage, organization, and computation that optimize analysis turnaround time and increasingly leverage the strength of Biota's proprietary subsurface DNA database, working towards a 'turn-key' computational pipeline implementation.
- Provide guidance in the development of current and future intellectual property and author top-tier journal publications to profile the cutting-edge science developed by Biota.
- Serve as an example for the company's values and behaviors as articulated in the Team Operating Agreement

Requirements:

- PhD level education (or MSc with 3-5 years relevant experience)
- 2 years of experience analyzing microbial ecology datasets using industry standard tools such as QIIME, MG-RAST, skbio, with co-authored publications (preferred).
- Scripting and data science experience in Python using sklearn and jupyter notebooks.
- Candidate should be highly proficient at performing analysis and scripting in R and Python with a track record of testing out new methods with minimal guidance.
- Skills related to primer design and variant detection in genomic DNA and RNA datasets. This includes experience with alignment and assembly tools like SAM tools and Bowtie.

Location: This position is based in our company's R&D center in San Diego or our Houston headquarter.

Compensation: Competitive base salary, bonus, stock options, and a benefits package including paid vacation, medical coverage, and telecommuting options.

Skills breakdown:

- Data analysis: 50%
- Software development: 30%
- Project management: 20%