Convergent Genomics Data Science Challenge

Primary Aims.

Through this challenge we are seeking to get to know you better as you demonstrate your critical thinking and creativity.

Deliverables.

1. A working pipeline implemented in python or R.
2. Data visualizations to help tell the story of your findings.
3. Communicate the rationale for your approach and results of your pipeline.
4. Written answers to each question section below.

The Challenge.
It is the mission of Convergent Genomics to bring clear and actionable insight to cancer patients and their physicians. One way you can support this mission is by leveraging the combination of clinical, laboratory, and sequencing data to create algorithms or classifiers that help diagnose disease and better understand risk, or aggressiveness of a patient's cancer. Through this challenge we are providing you an opportunity to demonstrate the raw power of your data ninja skills on a real world human cancer genomic and clinical feature dataset. Using the associated clear cell kidney cancer data sets provided, construct a program in python and/or R that:

1. Uses your choice of regression, clustering, or dimensionality reduction approaches to identify features underlying disease-associated risk.
2. Using best practices, create a classification algorithm that predicts risk.
3. Compare and contrast the optimal features identified in tasks 1 & 2.

Your choice of algorithms and graphical representations in addressing this challenge is intentionally left open ended. Report on those findings you consider most relevant to the goal.

Definitions.
Disease-associated risk: The clinical determination of risk is carried out through the combination of cancer stage (more information here), grade (more information here), overall survival in days following diagnosis, and vital status (alive/dead). These are presented for each patient ID in the `patient_data.tsv` file.

Inputs.

- Clinical data: `patient_data.csv`
- Tumor Mutation Sequencing data: `seq_data.csv`
- Tumor mRNA gene expression data: `mrna_data.csv`

Outputs.

- Working script(s).
- Relevant output data.
- Relevant descriptions accompanied by figures to support your findings.

Data Science Questions.

1. *What features of the data are most important for QC/QA?*
2. *Generally speaking, what are potential sources of ambiguity arising from your approach?*
3. *What other data might we collect to enhance risk quantification? What quantitative proof do you have?*
4. *Describe your approach to filing IP claims around your unique classification of risk?*
5. *How would you communicate your findings to a clinician?*

Questions?
Feel free to ask away: kevinphillips@convergentgenomics.com