Hello!

Thank you for your interest in working with us at the VUGENE company!

Here is a task for you to show your skills and knowledge in problem solving and computational skillset.

We are providing you with the *messy* sample sheet (where you can find sample ID (SID) and various covariates) and gene count matrix for RNAseq of the human brain. The datasets are taken from the study on alcohol use disorder (AUD). It would be very interesting to see what is differentially expressed when Control and AUD are compared.

The tasks are as follow:

1. Familiarise with information in the annoyingly *messy* sample sheet (thus realistic scenario). **Please tidy up and prepare this sample sheet** for it to be usable in the analysis. We would like it to be inspected and tidied up using code, as this approach aligns with our practice of ensuring reproducibility in our workflows. We would like to see what issues you noticed and what approach you used in overcoming the struggles to achieve the final *tidy* sample sheet.
2. Perform **Quality Control** of the data.
3. Perform data **normalization**, if needed.
4. **Differential expressed genes** (DEG) analysis (Control vs AUD). Detection of genes differentially expressed across sample groups using linear modeling and accounting for selected metadata covariates. Which variables should be used for linear modeling?
5. As the DEG can be very extensive in ways to visualise results, we only want to see the **volcano plot, marking the top 10 genes**.
6. Discussion: what else could one add to this analysis so it is the most current, efficient or effective. The goal of this exercise is to show how, after joining the team, could you help to improve our pipelines or process of work - such input is always very welcomed and encouraged!

Skills-wise we want to see your ability to work in R (but if you are more familiar with Python - go for it!), understanding of statistical analysis with additional emphasis on good practices for visualisations (at VUGENE we use ggplot a lot).

Afterwards, when we have a call with the team, we want you to present the results within the team, expressing your ways of tackling the tasks. We would also want to see your code during the presentation (with ability to have a look in more detal afterwards). Please, during the meeting with our team, use the opportunity to ask our bioinformaticians any questions you have.

Notes:

- We use R and Limma a lot: this reference might be useful to have a look at. Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies” 43: e47. https://doi.org/10.1093/nar/gkv007 .
- At the VUGENE we do a very wide range of analyses: metabolomics or proteomics, single cell or bulk RNA seq, methylation and epigenetics, as well as tackling synthetic biology or microbial sequencing projects. We do not expect a single person to join our team with expertise in all of the topics and skillsets, but we expect you to be curious, ask questions and learn how to efficiently work on these projects within the team.

Excited to meet you,
VUGENE team