**Bioinformatics pipeline #1 : PDL1-*as* as a prognostic marker in cancer**

Does the presence of alternative splicing in PDL1 (PDL1-*as*) correlate with improved prognosis in cancer patients? In the null model for this hypothesis, it is expected that the survival rate of PDL1-*as* positive cancer patients will not be statistically significantly different from the survival rate of cancer patients with no observed PDL1-*as* event. Survival is being defined as Progression Free Survival (PFS) in this model.

*Data source:* We need survival data, including progression/disease free survival (PFS, DFS), time from primary diagnosis to death (OS) **1**. This data is publicly available for 33 cancer types from TCGA. Expression data for PDL1 and PDL1-*as* isoform will be retrieved from the databases that have assumedly been used in the experiment outlined in the assignment (which found ~30% primary samples with PDL1-*as*), and that TCGA contains all these primary samples (if it does not, we expand search to ICGC). For multivariate analysis, we will also query TCGA for gender and age at diagnosis. The focus for this model is solely on primary tumor samples.

*Preprocessing and Statistics:* We identify the median DFS, PFS, and OS (in months) for all cancer types being studied. The median survival time is tested using a likelihood ratio test, with the non PDL1-*as* patients as the control set in each cancer type. A univariate analysis can be done by building Kaplan-Meier survival curves for PDL1-*as* and non PDL1-*as****\**** groups, or alternatively by statistically testing the difference between these two groups using logrank test**2**. However, to account for impact of other factors (such as gender, age, other mutation status’ of PDL1), we can also carry out a multivariate analysis. The commonly used Cox test on survival time can be used to identify the ‘hazard ratio’ of PDL1-*as* positive status**2**. If PDL1-*as*’s hazard ratio is < -1, then PDL1-*as* positive patients likely have a positive prognosis. The robustness of our belief in the prognostic value of PDL1-*as* status is evident in the rigorous testing through different statistical approaches listed here (univariate and multivariate models). If PDL1-*as* is a significant predictor, it will be robust to addition of other explanatory covariates in our model.

**\****For the sake of brevity we use a simple 2-group classification based on PDL1-as status. Ideally we should also account for other inactivating mutations in PDL1, contingent on access to patient-specific somatic mutation data from TCGA. We then combine the mutation status with the expression measurements of PDL1 (not PDL1-as) as a proxy for the functional state of PDL1 in all patients (need differential expression analysis to identify which PDL1 mutation sets have a corollary differential expression versus ‘normal, functional PDL1’ samples). In that case, we will ultimately be comparing multiple groups (PDL1-as, PDL1-mut1, PDL1-mut2 etc.), and will need to correct all results from our listed statistical results for multiple testing (ex. Bonferroni, if the groups are disjoint).*

**Bioinformatics pipeline #2**

Is there a positive selection for the PDL family in the human lineage? In the null model for this hypothesis, it is expected that the rate of accrual of non-synonymous (amino-acid-altering) base changes in gene members of PDL family is no different from the rate of accrual of synonymous base changes since divergence of modern humans from archaic humans (represented by the archaic human in this study, and sequenced Neanderthal genomes).

*Data source*: Identify PDL1 paralogues using GenBank

*Statistics*: Identify orthologous sequence pairs using McDonald Kreitman method**4** ????. Identify areas under selection pressure using Tajima’s relative rate test. Estimate number of non-synonymous substitutions (Na) and number of synonymous substitutions (Nb) for every sequence under consideration using the modified Nei-Gojobori method. Calculate D (= Na - Ns) for all between-group and within-group comparisons, and use a statistical test to measure if the D is significantly different (i) between the archaic and modern human samples, and (ii) within the respective specie-al group (archaic, modern) (to account for genetic drift). Positive selection for a loci with be indicated by a positive, high value of D, and a multiple-testing corrected p-value for statistically significance difference in D in archaic and modern humans.**3**

*Future directions/wishful thinking*: Does PDL1-*as* have a basis in patient demographics? Is a poorer prognosis in the presence of PDL1-*as* correlated with a more recent migration of the respective ethnic group from Africa? Is PDL1-*as* isoform also seen in relapse patients, or patients with metastatic cancer?

Works Cited

1. <http://www.nature.com/articles/srep02246>
2. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2394368/
3. <http://www.nature.com/nature/journal/v413/n6855/full/413514a0.html>
4. http://genome.cshlp.org/content/24/6/885.full