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

Does the presence of alternative splicing in PD-L1 (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.

Bioinformatics pipeline #2: Is there a positive selection for mutations in PD-L1 in humans?

Has PD-L1 acquired more frequent, advantageous amino-acid changing mutations over time than it should under normal genetic drift? In the null model for this hypothesis, it is expected that the rate of accrual of non-synonymous (amino-acid-altering) base changes in PD-L1 is no different from the rate of accrual of synonymous base changes, since divergence of modern humans from archaic humans (composed of the archaic human in this study, and Neanderthal genomes**).[5]

Data source: Get modern day humans' somatic mutation data from 1000 Genomes, expression data from Guevadis Projects, and ancient hominin genomes and transcriptomes from previously published papers[7]. We can retrieve RNA-Seq and somatic mutation data from the archaic human mentioned in this study, and for Neanderthals[7] (use UCSC or Ensembl). Identify polymorphisms specific to modern humans in the PD-L1 gene, and polymorphisms present in early hominins, using the somatic mutation data at hand.

Statistics: We first estimate the number of non-synonymous substitutions (D_n) and number of synonymous substitutions (D_s) separately for PD-L1 and for a shared pseudogene (since pseudogenes generally evolve neutrally with $D_n/D_s \sim 1$, lacking strong selection for/against any new mutations[5]). Several pseudogenes shared in humans and Neanderthals in olfactory receptors have been identified, and can be used as test sets for neutral genes[8]. Calculate D ($= D_s - D_n$) for all between-group and within-group comparisons, and fit models to test if D is significantly different for PD-L1 versus 'neutral genes', in the modern human genomes versus archaic humans. We can use the PAML package to fit models of positive selection. Here, a subset of codons (namely, PD-L1 codons) will be expected to evolve with Dn/Ds > 1 in our alternate hypothesis, whereas the null model will assume there is no positive selection for any codons being evaluated. We then use a likelihood ratio test to evaluate if our data fit the positive selection model better than the null model.[3,5] This approach can be made robust by iteratively selecting different pools of genes as our 'neutral genes' and carrying out the PAML model fit approach.

Notes on robustness

Pipeline 1:

The robustness of our belief in the prognostic value of PDL1-*as* status is evident in the rigorous testing through different statistical approaches listed earlier in the discussion (univariate and multivariate models). In additional, 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).

Pipeline 2:

20% of the Neanderthal genome is shared with modern humans[7]. We can use documented genes undergoing positive selection over human evolution (BRCA1, BRCA2) as our positive replicates (in place of PD-L1) and see if a likelihood ratio test using this model still upholds our alternate hypothesis of positive selection in the 'preferred genes' set (PD-L1, BRCA1/2). All the genes (neutral and PD-L1) considered in our model can also be evaluated using Fay and Wu's H test – we would expect the derived allele frequencies in PD-L1 to be higher than that in the neutral genes, since that would indicate acquisition of more mutations than by random chance in PD-L1. An LRH test (long range haplotype test) can also be used to test for gene level positive selection, by comparing disequilibrium across haplotype blocks in the PD-L1 gene, versus a neutral gene. We would expect fewer haplotype blocks with increased similarity than by random chance, in the neutral gene, whereas if positive selection did indeed happen for PD-L1, more haploytype blocks in that gene will be 'preserved' across the archaic and modern humans.

Extra notes

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, and if so, how well does this correlate with survival? A similar approach as Pipeline#2 can be applied to modern healthy humans versus modern humans with cancers that exhibit PD-L1 mutations. For the latter, we will need to expand our data sources to include TCGA samples, and require somatic mutation data (which is protected access) and normalized RNA-Seq expression data (public access).

***I am assuming that the PDL-1 homolog is known in the Nenaderthal genome, as the question states the H. Senescesaurus, which has the PD-L1 gene, is the most long-lived human ancestor known (so presumably was also there around/before the Neanderthals died out as our early ancestors).*

References

1. Yang, P.; Du, C. W.; Kwan, M.; Liang, S. X.; Zhang, G. J. (2013-07-22). "The impact of p53 in predicting clinical outcome of breast cancer patients with visceral metastasis". *Scientific Reports* **3**. doi:10.1038/srep02246. PMC 3718193. PMID 23873310.
2. Bradburn, M J; Clark, T G; Love, S B; Altman, D G (2003-08-04). "Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods". *British Journal of Cancer* **89** (3): 431–436. doi:10.1038/sj.bjc.6601119. ISSN 0007-0920. PMC 2394368. PMID 12888808.
3. Johnson, Matthew E.; Viggiano, Luigi; Bailey, Jeffrey A.; Abdul-Rauf, Munah; Goodwin, Graham; Rocchi, Mariano; Eichler, Evan E. (2001-10-04). "Positive selection of a gene family during the emergence of humans and African apes". *Nature* **413** (6855): 514–519. doi:10.1038/35097067. ISSN 0028-0836.
4. Enard, David; Messer, Philipp W.; Petrov, Dmitri A. (2014-06-01). "Genome-wide signals of positive selection in human evolution". *Genome Research* **24** (6): 885–895. doi:10.1101/gr.164822.113. ISSN 1088-9051. PMC 4032853. PMID 24619126.
5. Dannemann, Michael; Andrés, Aida M.; Kelso, Janet (2016-01-07). "Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors". *American Journal of Human Genetics* **98** (1): 22–33. doi:10.1016/j.ajhg.2015.11.015. ISSN 1537-6605. PMC 4716682. PMID 26748514.
6. Lou, Dianne I; McBee, Ross M; Le, Uyen Q; Stone, Anne C; Wilkerson, Gregory K; Demogines, Ann M; Sawyer, Sara L (2014-07-11). "Rapid evolution of BRCA1 and BRCA2 in humans and other primates". *BMC Evolutionary Biology* **14** (1). doi:10.1186/1471-2148-14-155. PMC 4106182. PMID 25011685.
7. Gokhman, David; Lavi, Eitan; Prüfer, Kay; Fraga, Mario F.; Riancho, José A.; Kelso, Janet; Pääbo, Svante; Meshorer, Eran; Carmel, Liran (2014-04-17). "Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan". *Science*: 1250368. doi:10.1126/science.1250368. ISSN 0036-8075. PMID 24786081.
8. Curnoe, Darren. "Making sense of our evolution". *The Conversation*. Retrieved 2016-03-22.