Does tumour purity have an effect on the ability to detect allele-specific expression (ASE)?

Introduction

Allele-specific expression (ASE) refers to the differential or unequal expression of gene copies. While differences in gene expression between alleles have previously been investigated in the context of epigenetics and imprinting, this form of allelic expression imbalance occurs due to cis-acting, heritable genetic variation among autosomal genes[1].

ASE has also been shown to play a role in cancer development and is frequently observed in tumour samples. Somatic mutations in the cis-regulatory elements which regulate ASE genes may be driver mutations which contribute to cancer[2]. A loss-of-function mutation in a recessive tumour suppressor gene can lead to cancer when the more highly expressed allele is affected[3,4]. In addition, ASE can also cause the overexpression of genes with oncogenic mutations, which can disrupt homeostasis and promote the initiation of cancer[5].

This project aims to investigate whether the tumour purity, or percentage of cancer cells present in the tumour tissue, has an effect on estimating ASE. Through analysing sequencing data from pancreatic tissue samples using R statistical software, we investigated whether adjusting count values to account for tumour purity affected our ability to detect ASE. The data used for analysis had previously been run through cisASE, a likelihood-based method used to identify cis-regulated ASE from sequencing data (Liu et al., 2016). Generally, cisASE has been used with matched DNA-RNA sequencing data, however it is considered applicable for datasets without DNA-seq, and thus another aim of our project was to investigate whether ASE could be detected in genes using only RNA-seq data.

Limitations and assumptions

There are a number of caveats which we employed and limitations which we recognise to our project. Firstly, we are assuming that the normal sample only contains the reference allele, and that the alternate allele is only found in the tumour sample. We also assumed that if a gene was not included in the GTEx data of genes normally expressed in pancreatic tissue, it was also not expressed in the non-tumour tissue of our samples.

In addition, for our initial investigation, we assumed that there was no alignment bias in favour of the reference over the alternative allele which could be contributing to differences rather than tumour purity. When the alternative count was found to be higher than the reference count, we assumed that these were true candidate genes showing ASE because even with alignment bias in favour of the reference allele and contamination of the normal sample the reference count was higher.

Materials and methods

The pancreatic tissue samples used for analysis in this project came from the cancer genome atlas (TCGA) dataset[7]. The software tool cisASE was used to identify candidate ASE genes from matched DNA-RNA as well as RNA sequencing only data[6]. The R statistical software (version 2022.02.1) was used for data analysis.

The output from cisASE was initially analysed to compare between the matched DNA-RNA results and the RNA-only results. For the gene-level cisASE results we applied the recommended log-likelihood ratio (LLR) (>0.82) and heterogeneity p-value (>0.05), to remove splicing errors, thresholds to identify candidate ASE genes. Genes that showed copy number variation (CNV) were also removed to control for bias.

We used median TPM gene expression data for normal pancreatic tissue from GTEx[9] to identify which genes are usually and not usually expressed in this tissue type. A number of reference databases were used to investigate gene functions and expression – GeneCards[8], GTEx[9], NCBI gene database[10], and COSMIC[11].

Single nucleotide variants (SNVs) cisASE results were used to identify candidate ASE SNVs using both the default LLR threshold (0.82) as well as the sample specific LLR cut-offs (Table 1) determined through 2000 rounds of simulations.

Sample	Analysis	LLR Threshold (0.05 significance level)
111	RNA only - Gene	0.70
111	Matched RNA\DNA - Gene	0.72
111	Matched RNA\DNA - SNV	0.84
111	RNA only - SNV	0.93
131	RNA only - Gene	0.88
131	Matched RNA\DNA - Gene	0.82
131	Matched RNA\DNA - SNV	0.87
131	RNA only - SNV	0.91
161	RNA only - Gene	0.89
161	Matched RNA\DNA - Gene	0.83
161	Matched RNA\DNA - SNV	0.89
161	RNA only - SNV	0.90

Table 1: New LLR thresholds applied to the data specific to sample and analysis type

We adjusted the RNA count values based on tumour purity in order to see whether this had an effect on the candidate ASE genes identified through our analysis. Sample 111 had a purity level of 39%, Sample 131 had a purity of 72%, and Sample 161 had a purity of 19%. The adjusted RNA reference count value was obtained through a series of calculations incorporating tumour purity. First, the normal sample fraction was found (1 – tumour purity). The total counts were also calculated by summing the Ref and Alt counts. Then, the total number of reads in the sample attributed to the normal sample was found by multiplying the normal sample fraction by the total count value. Finally, the adjusted Ref count value was calculated by subtracting the total reads attributed to the normal sample from the Ref count. This final value corresponded to the total number of reference reads attribute to the tumour sample.

The negative adjusted reference counts were assumed to be candidate ASE genes and were separated from positive adjusted count values. A binomial test was performed using the positive subset of adjusted reference count to find candidate ASEs in this group. The p-value was also adjusted using the false discovery method (FDR) for multiple test correction.

Results

Candidate ASE genes:

Table 2 shows a comparison of the number of genes per sample before and after filtering for significant LLR and p-value thresholds, as well as removal of genes showing CNV, resulting in a significant decrease in the number of candidate ASE genes. One of our samples, sample 111, is hypermutated and therefore has a larger number of SNVs compared to the other two samples (Table 2). This resulted in a larger number of genes assessed for ASE for this sample. The number of ASE candidate genes were similar when we compared the results from the matched DNA-RNA to the RNA-only candidates.

	Sample 111		**Sample 131**		**Sample 161**	
	Matched	*RNA only*	*Matched*	*RNA only*	*Matched*	*RNA only*
Before filtering	7841	7835	551	551	404	404
LLR > 0.82	4123	4269	262	328	185	164
p-value >=0.05	1047	1036	7	15	4	4
CNV = 0	987	989	7	11	4	4

Table 2: Comparison of candidate ASE genes per sample for matched and RNA only sequencing data.

Identification of genes that are not usually expressed in the pancreas:

We identified 5 candidate ASE genes (Table 3) not usually expressed in pancreatic tissues, using both the gene-level and SNV-Level cisASE results. These were Transmembrane Serine Protease 15 (*TMPRSS15*), H2B Clustered Histone 12-like (*H2BC12L*), Long Intergenic Non-Protein Coding RNA 1618 (*LINC01618*), Acetylserotonin O-methyltransferase-like (*ASTML*), and A-Kinase Anchoring Protein 17A (*AKAP17A*). After further analysis using the reference databases GeneCards[8], GTEx[9], NCBI gene database[10], and COSMIC[11], 2 of the 5 genes were found to have a potential association with colorectal and/or pancreatic cancer[12,13]. However, there are also other biological and technical factors which could indicate genes not usually expressed in the pancreas, such as blood contamination.

Gene ID	**Analysis type**	**Gene name**	**Function**
ENSG00000154646	Matched, RNA only	TMPRSS15 - transmembrane serine protease 15	• Enzyme which activates pancreatic proteolytic proenzymes. • Function is in pancreas, but expression restricted towards duodenum.
ENSG00000234289	Matched, RNA only	H2BC12L – H2B clustered histone 12 like	• Enables DNA binding
ENSG00000250302	Matched	LINC01618 – Long intergenic non-	• Previously associated with colorectal cancer[12]

		protein coding RNA 1618	
ENSG00000169093	Matched, RNA only, SNV	ASTML – Acetylserotonin O-methyltransferase like	• Pyrophosphatase • Methyltransferase • Previously associated with colorectal/pancreatic cancers[13]
ENSG00000197976	RNA only, SNV	AKAP17A – A-kinase anchoring protein 17A	• Part of spliceosome complex

Table 3: Candidate ASE genes and SNVs which are also not usually expressed in pancreatic tissue

Re-analysis using specific LLR thresholds:
When we re-analysed this data using LLR thresholds specific to each sample and analysis type, we found no difference in the number of candidate ASE genes identified (Table 4) indicating that the default LLR threshold is sufficient to call candidate ASE genes. We also confirmed that the genes identified using both thresholds were the same.

	Number of unique candidate ASE genes	
	LLR > 0.82	**Specific LLR threshold**
Matched gene	10	10
RNA only gene	9	9
Matched SNV	12	12
RNA only SNV	10	10

Table 4: Comparison of unique candidate ASE genes found for each analysis type with the old and new LLR thresholds applied

Comparison of methods used to identify ASE SNVs
The Venn diagrams in Figure 1 below illustrate the findings of comparisons between different methods of identifying candidate ASE SNVs. Candidate ASEs were identified using the cisASE results with the LLR threshold of > 0.82 applied, as well as by running a binomial test on reference counts which were both adjusted and unadjusted to account for tumour purity to investigate whether this affected results.

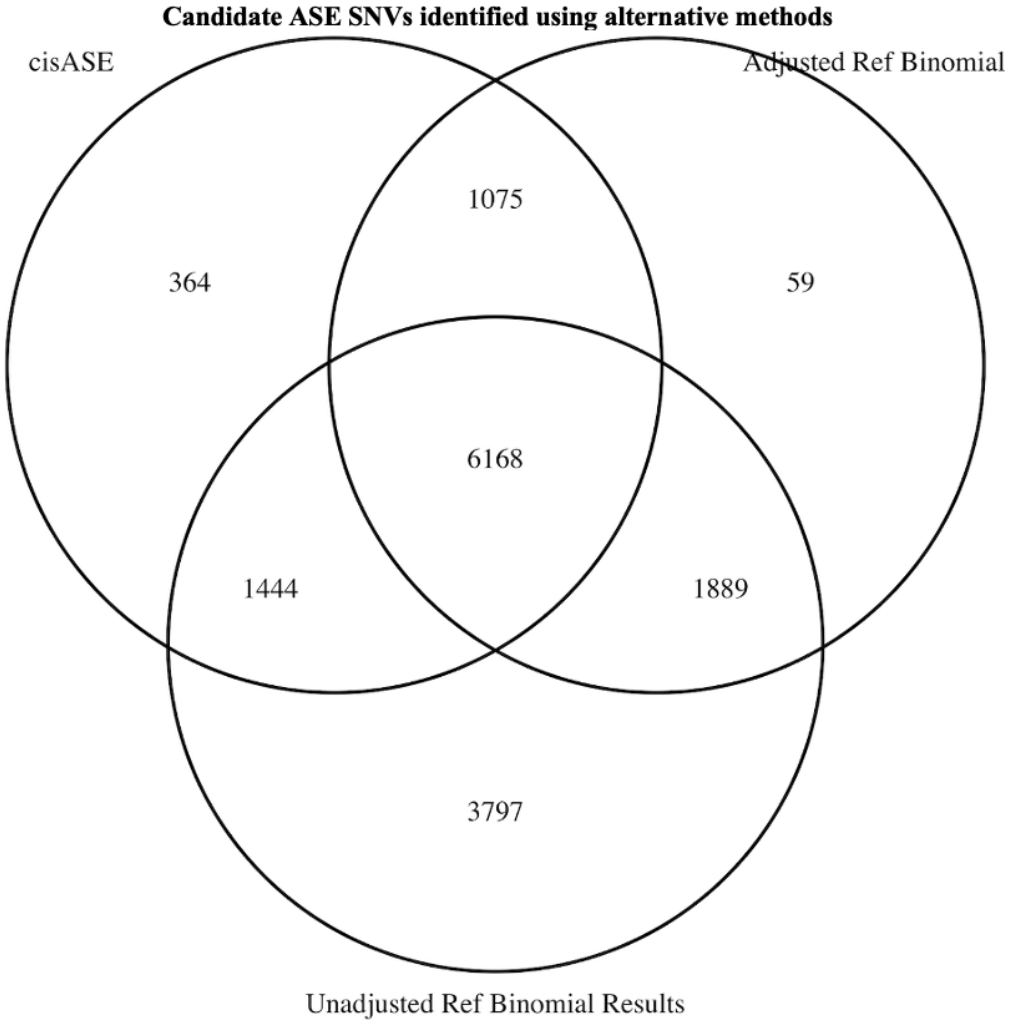

Figure 1: Venn diagram showing comparison of candidate ASE SNVs identified using **(i)** cisASE with LLR threshold, **(ii)** binomial test with unadjusted reference count and **(iii)** binomial test with adjusted reference counts

Discussion

Findings on using cisASE:

The potential of software such as cisASE for applications in detecting genes which show allele-specific expression is huge, particularly in developing our understanding of the mechanisms of cancer initiation and progression. This project explored the process of identifying candidate ASE genes from cisASE data by applying different significance thresholds and investigated the difference in candidate calls when using matched DNA-RNA data compared to RNA-only data.

The similar results obtained using RNA-only data and those using matched DNA-RNA data (Table 2) indicate that cisASE can detect ASE in genes from using only RNA-seq data. Considering the current costs and storage constraints involved with sequencing large amounts of data, it is hugely advantageous that these tests to identify genes showing ASE

can be performed using only RNA-seq data, without the need for a matched DNA-seq sample too.

We found that identifying candidate ASE genes using a standard LLR cut-off of 0.82 yielded the same results as the alternative method of applying individual specific LLR cut-offs depending on the sample and analysis type. This finding indicates that the default LLR threshold of > 0.82 is sufficient to identify candidate ASE genes from the data, without the need to parse out the sample-specific thresholds from the cisASE result files.

Impact of adjusting SNV based on tumour purity:
A comparison of methods to identify genes showing ASE (Figure 1), namely candidates found using an LLR threshold applied to cisASE results and those identified using binomial tests with both adjusted and unadjusted reference counts yielded several important observations. Firstly, it is clear that neither binomial method is picking up all of the candidates found using cisASE, and there are additional candidates detected by the binomial tests which cisASE doesn't call. This may be explained by cisASE using models to account for alignment bias[6], while the binomial test is a much simpler model which has not accounted for this bias. It is also significant that the total number of candidates identified by a binomial test that were not called by cisASE, or possible false positives, were much lower for the results with adjusted ref counts (1948) when compared with the unadjusted test (5685). This perhaps implies that our adjustments for tumour purity have improved the ability to detect true candidate ASE genes with a binomial test.

There are a number of possibilities which could explain the subsets of genes called by cisASE and unadjusted or adjusted binomial test only. There was a larger overlap between those called by cisASE and unadjusted test only (1444) than by cisASE and adjusted only (1075), and it could be the case that the former candidates are in fact false positives if cisASE does not specifically account for tumour purity. The binomial test adjusting for purity may be detecting less false cisASE candidates, in which case, accounting for this when identifying ASE does have an effect on results. Alternatively, the subset of candidates called by cisASE and adjusted binomial test could be showing that the cisASE model itself does unintentionally account for differences in tumour purity to some extent. However, this could also be explained if adjusting the reference count based on tumour purity has, by chance, captured the cisASE adjustment for alignment bias for some of the SNVs.

Future work
In conclusion, our analysis has found several results and possibilities which require further investigation to provide more conclusive answers. Additional research is required to account for confounding variables which could affect our results. Accounting for alignment bias in this sequencing data in particular is vital to see whether this has an effect and subsequently apply new methods to adjust for it if it does. This could be performed by considering genome mappability, or the ability of sequence reads to uniquely map to individual regions of the genome[14].

Running these tests and adjustments on larger TCGA datasets with more samples would also provide important insights into the accuracy of these findings. Currently, we cannot rule out that the candidates detected by our adjusted binomial method are either true or false ASE

SNVs. It would be useful to repeat the same method on simulated data, or on data that has previously been experimentally validated, to investigate the accuracy of the model. Our findings suggest that tumour purity may have an effect on the ability to detect ASE, however additional work and refinement of our model is necessary to confirm this.

References

1. Knight, J.C. (2004). Allele-specific gene expression uncovered. *Trends in Genetics*, 20(3), pp.113–116. doi:10.1016/j.tig.2004.01.001.
2. Liu Z, Dong X, Li Y. (2018). A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *Front Genet*. 2018;9:570. doi:10.3389/fgene.2018.00570
3. Rhee, J.-K., Lee, S., Park, W.-Y., Kim, Y.-H. and Kim, T.-M. (2017). Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. *Scientific Reports*, 7(1). doi:10.1038/s41598-017-01966-z.

4. Clayton, E.A., Khalid, S., Ban, D., Wang, L., Jordan, I.K. and McDonald, J.F. (2020). Tumor suppressor genes and allele-specific expression: mechanisms and significance. *Oncotarget*, 11(4), pp.462–479. doi:10.18632/oncotarget.27468.

5. Bielski, C.M., Donoghue, M.T.A., Gadiya, M., Hanrahan, A.J., Won, H.H., Chang, M.T., Jonsson, P., Penson, A.V., Gorelick, A., Harris, C., Schram, A.M., Syed, A., Zehir, A., Chapman, P.B., Hyman, D.M., Solit, D.B., Shannon, K., Chandarlapaty, S., Berger, M.F. and Taylor, B.S. (2018). Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. *Cancer Cell*, 34(5), pp.852-862.e4. doi:10.1016/j.ccell.2018.10.003.

6. Liu, Z., Gui, T., Wang, Z., Li, H., Fu, Y., Dong, X. and Li, Y. (2016). cisASE: a likelihood-based method for detecting putativecis-regulated allele-specific expression in RNA sequencing data. *Bioinformatics*, 32(21), pp.3291–3297. doi:10.1093/bioinformatics/btw416.

7. TCGA: The Cancer Genome Atlas - https://www.cancer.gov/tcga

8. GeneCards: the human gene database - www.genecards.org
   Safran M, Rosen N, Twik M, BarShir R, Iny Stein T, Dahary D, Fishilevich S, and Lancet D. *The GeneCards Suite*Chapter, *Practical Guide to Life Science Databases* (2022), pp. 27-56
9. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J. and Fleming, J. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), pp.580–585. doi:10.1038/ng.2653.

10. Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y. and Tse, T. (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Research*. doi:10.1093/nar/gkab1112.

11. COSMIC: the Catalogue of Somatic Mutations in Cancer
cancer.sanger.ac.uk
Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S.C., Kok, C.Y., Noble, K., Ponting, L., Ramshaw, C.C., Rye, C.E. and Speedy, H.E. (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1), pp.D941–D947. doi:10.1093/nar/gky1015.

12. Wang, X., Liu, Z., Tong, H., Peng, H., Xian, Z., Li, L., Hu, B. and Xie, S. (2019). Linc01194 acts as an oncogene in colorectal carcinoma and is associated with poor survival outcome. *Cancer Management and Research*, Volume 11, pp.2349–2362. doi:10.2147/cmar.s189189.

13. Bi, J., Huang, Y. and Liu, Y. (2019). Effect of NOP2 knockdown on colon cancer cell proliferation, migration, and invasion. *Translational Cancer Research*, 8(6), pp.2274–2283. doi:10.21037/tcr.2019.09.46.

14. Karimzadeh, M., Ernst, C., Kundaje, A. and Hoffman, M.M. (2018). Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research*, 46(20), p.e120. doi:10.1093/nar/gky677.