Pangenomes: A better reference for a more diverse dataset?

Carlos Avendaño
Molecular Cellular Biology Program
University of Washington
Mohammadi Lab

Abstract

New development of tools to work with graph references hopes to eliminate bias. Pangenome may better represent a diverse cohort. A newly released tool, RPVG, allows us to perform haplotype expression, using pangenome graphs. This project assesses if RPVG can produce quality ASE data, and if pangenome references are an improvement over already existing methods.

1 Introduction

1.1 ASE

Allele Specific Expression or ASE is a process where one allele is over or under-expressed relative to the others. While ASE can occur with any ploidy, it is easiest to study and observe in humans as they are diploid organisms, getting a chromosome from each parent. RNA sequencing can be used to study this phenomenon and determine expression levels for each haplotype. [1]

RNA reads are mapped to a reference, where reads are counted whether they map to the reference or alternative haplotype with a tool such as ASEReadCounter or phASER, depending if ASE analysis is population or variant level. [1] To run phASER, data must be phased first, with a tool such as Eagle2, to determine which haplotype a SNPs originated from to. [7] Allelic expression can be modeled using a binomial test, with most genes expressed equally.

ASE data is often overexpressed, thus additional methods have been developed to help clean up the results. Genome Analysis Toolkit (GATK) can remove duplicated reads and ANalysis of Expression VAriation (ANEVA) can remove outliers while keeping truly overexpressed genes. [8] [9]

Locating which heterozygous sites and mutations are respossible for this imbalance in gene expression can help with understanding some of the causes of rare and common diseases in people. [10]

1.2 Pangenomes

One current issue in the field of ASE and human genomics as a whole is reference bias, where the patient or sample genomically distant enough from the reference used. Reads may not map properly to the reference. This is especially prevalent in people who are not of European ancestry, as the majority of the samples, references and research into the human genome is based off of humans of European ancestry. [4]

Pangenomics is a potential solution to the issue of reference bias, where graph references can be built from multiple distinct references to account for genomic diversity. Figure 1 shows an example of a pangenome graph. There are regions of the pangenome that are shared between all references (1) used to create the graph. There are also alternative pathways that are unique to a given reference (2 and 3). Pangenomes, carry the quality of multiple references, but are much larger and are, more computationally intensive to use.

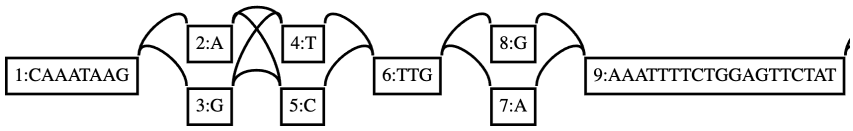


Figure 1: Example Graph Showing alternative alighment paths. Figure was generated using VG Viz from the VG Tools Package, using one of their example pangenomes.

1.3 VG Tools

VG Tools is a package that can be used to build pangenome graphs from scratch with fasta references, modify existing pangenome graphs and map RNA sequencing reads to these constructed graphs. Pangenomes are difficult to work with, being very computationally intensive, VG Tools runs efficiently as it converts graphs and alignment files into proprietary index files.

The VG MPMAP tool, written specifically for VG Tools, can map reads to a pangenome reference. It uses a 'seed–cluster–extend' paradigm, where it first maps smaller chuncks of a read to a reference, to check for a possible match before extending and checking the rest of the read. Pangenome graph references are much larger than regular references, with multiple paths a read could map to. Breaking up the reads into smaller groups has been shown to be more efficient, as the computationally intensive full mapping step is only done if the smaller "seed" matches a given region first, rather than checking the entire length of the read. The VG MPMAP tool produces a multipath alignment file as its output. [3] [6]

1.4 RPVG

The VG Team recently released a new tool, RPVG, which claims to be able to perform haplotype expression using pangenome graphs and pantranscriptomes generated from VG Tools. Pantransciptomes are generated with pangenome construction in VG tools.

Reads are mapped to a pantranscriptome using a multipath alignment and counted. Partial alignments are counted as well. To improve computational efficiency, reads that map to shared paths are clustered, and clusters are worked in parallel, with new graphs constructed for each cluster.

In the VG Team's RPVG paper, they claim that the tool can be used to generate ASE data, and that it is more effective than other methods typically used to generate ASE data. [11]

2 Rationale

The VG Team tested their tool RPVG, on only a few samples. They compared the output of their tool with simulated reads as a comparison against other read counting tools: RSEM, Salmon and Kallisto. They found that using their tools, and a pangenome reference, they could map more reads, and generate more counts, compared to other tools. The VG Team did not do any ASE analysis, despite claiming their tool can be used to generate this type of data.

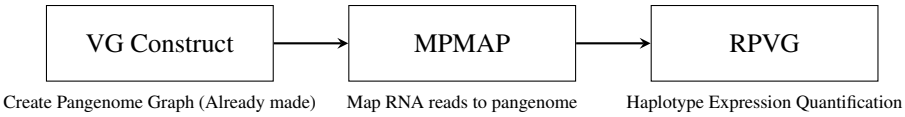
The purpose of this project is to determine if RPVG can generate quality ASE data, and if using a Pangenome reference produces more accurate data than traditional methods and using a standard reference.

3 Methods

3.1 Running MPMAP and RPVG

Mapping to pangenomes is computationally intensive, all work is done on Seattle Children's Research Institute's High-Performance Computing Cluster.

In order to scale up and help properly manage memory and CPU allocation, a Nextflow pipeline was written to run the VG Tools and RPVG in parallel for multiple samples.



Pangenomes constructed were constrcuted by the VG Team, using samples from the 1000 Genomes Project.The same pangenomes and pantranscriptomes used in their paper were used in this analysis, the pangenome used was created from all the european samples in the 1000 genomes database.

The VG Team published their Pangenomes and Pantranscriptomes on UC Santa Cruz's server

`http://cgl.gi.ucsc.edu/data/vgrna/pantranscriptomes/`

The VG Team as of writing this report has updated the format of their indexes for their pangenomes, thus an apptainer container, was built to run the older version of VG Tools. Version 1.38.0 of VG Tools was used, as it was the same version used in the RPVG paper.

Example MPMAP command:

```
vg mpmap -n rna -t ${task.cpus} -x ${XG} -g ${GCSA} -d ${DIST}
-f ${R1} -f ${R2} > ${GAMP}
```

".xg": Spliced pangenome graph

".gcsa": GCSA index of pangenome graph

".gcsa.lcp": LCP array for gcsa

".dist": Distance index of spliced pangenome graph

".gbwt": Pantranscriptome

".gbwt.ri": r-index of pantranscriptome

".txt.gz": Transcript and haplotype information

"R1 and R2": Paired end RNA sequencing reads

".gamp": compressed allignment file

According to the VG Team, RPVG, does not work properly in a containerized environment, but the team provided a pre compiled binary, which was used to run RPVG. Results from MPMAP are fed into RPVG.

Example RPVG command:

```
rpvg --use-allelic-mapq ${task.cpus} -g ${XG} -p ${GWBT} -a ${GAMP}
-o ${OUTPUT} -i haplotype-transcripts -f ${TRANSCRIPT}
```

4 Results

4.1 Accessing RPVG with Kallisto

The VG Team published their ASE analysis, but only preformed their full analysis on one sample: NA12878

`https://zenodo.org/records/7234454`

Unlike standard ASE analysis, the VG team compared Transcripts per million (TPM) rather than counts per gene. While unorthodox, other tools such as Kallisto output their results as TPM, although Kallisto is not typically used for ASE analysis. RPVG produces results as either TPM or counts per

transcript. Results from the VG team report were compared to synthesized data to assess the quality of their data. While their report compares RPVG to other popular RNA quantification tools (Kallisto, Salmon, RSEM), they compare each tool against simulated data rather than against one another.

To STARt this report, expands on the VG Team's comparision, by directly comparing the outputs of RPVG and Kallisto. Kallisto was chosen because it quantifies read counts on transcripts level like RPVG and also produces results as TPM, to follow the same methods the VG team used. The same sample, NA12878, from the same dataset from the RPVG paper was used in this comparison.

RPVG

Statistic	Value
count	$1.846\,557 \times 10^8$
mean	0.362 227 3
std	20.048 58
min	0.000 000
25%	0.000 000
50%	0.000 020 748
75%	0.000 230 361
max	21 445.45

Kallisto

Statistic	Value
count	$1.846\,557 \times 10^8$
mean	0.008 625 856
std	3.480 867
min	0.000 000
25%	0.000 000
50%	0.000 000
75%	0.000 000
max	17 471.70

Table 1: RPVG and Kallisto TPM Summary Statistics

Outputs of RPVG and Kallisto were compared, filtered for transcripts that mapped to both tools, Table 1. RPVG produced a significantly higher mean TPM than Kallisto with, a 42-fold increase in average TPM. This matches the VG Team's claim that RPVG can map more reads, using a pangenome reference rather than a conventional fasta reference. While RPVG mapped more reads than Kallisto, this does not necessarily, translate to higher quality data, as the standard deviation of 20 for RPVG. This is much higher than Kallisto's standard deviation of 3.5. RPVG produces results that are much more dispersed, than conventional methods.

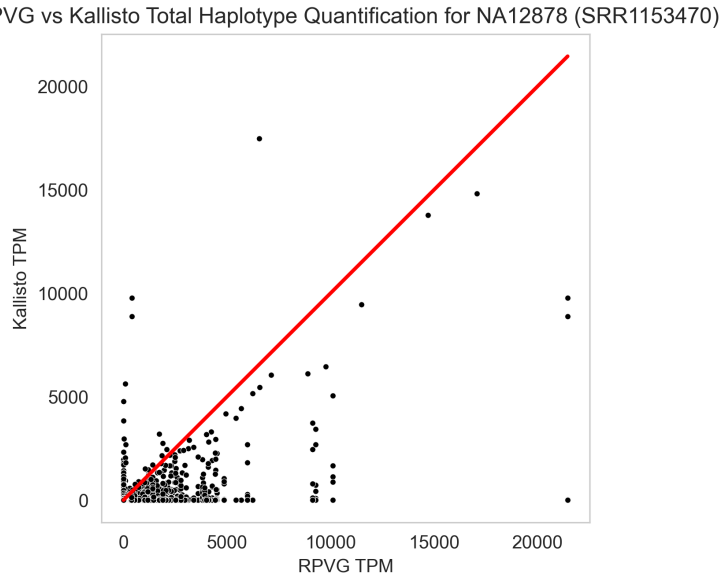


Figure 2: Kallisto and RPVG Transcripts Per Million (TPM) for NA12878, the sample used in the paper. A line with a slope of 1 shows where TPM for both methods are equal. A filter to remove transcripts with less than 50 TPM for where total haplotype count was less than 50, was applied to make visualization more clear.

Figure 2 shows that while RPVG mapped more reads for the majority of transcripts, compared to the output of Kallisto, the increased number of read counts per transcript, is not consistently increasing. Some transcripts are more expressed than others relative to the results from Kallisto. A more the linear trend would be expected if RPVG was producing better results than Kallisto, with all transcripts showing a similar increase in counts. RVPG was able to map many more transcripts than Kallisto, which may be in part due to using a Pangenome reference which is more inclusive than a standard reference. RPVG also has a more lenient mapping algorithm, as RPVG maps partial transcripts, and includes it in the read count.

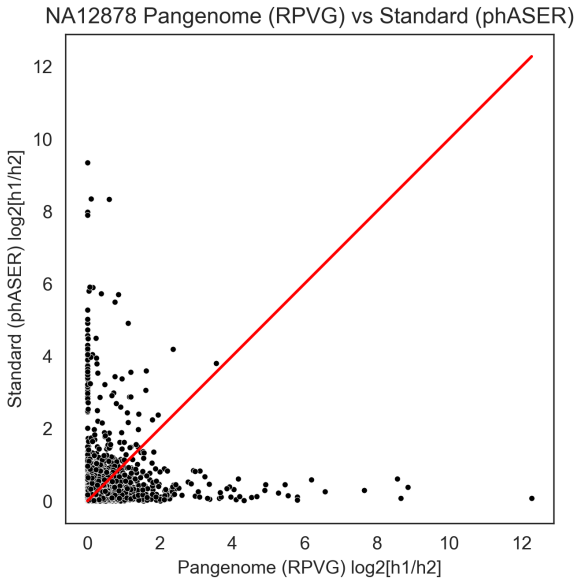


Figure 3: Pangenome (RPVG) vs Standard (phASER) log 2 [Hap 1/ Hap 2] for NA12878. A filter to remove total gene counts of less than 50, was applied to make visualization more clear.

RPVG was compared against, standard ASE data generation methods (Alignment with STAR [5], counting and read counting done with phASER [2]). The comparison was done by taking the natural log of the ratio of haplotype 1 divided by haplotype 2, for each gene, for RPVG and Standard ASE methods for the same sample. This allowed a comparison of the level of expression of each haplotype from both methods. Figure 3 shows no clear linear relationship between the two methods, while RPVG and MPMAP may produce more counts, it is not able to consistently discern which, genes are overexpressed, a somewhat linear relationship would be expected if RPVG could properly determine the same level of expression as standard methods.

In order to compare RPVG results which are on a transcript level, to standard ASE analysis methods, RPVG results needed to be converted to gene level. This was done by mapping back RPVG transcript outputs back to the original annotation file which contained a key of genes and transcripts. Transcrips were summed, for genes that contained multiple transcripts. The comparison of RPVG and Standard methods were done with counts per gene.

4.2 Accessing RPVG with MAGE dataset

MAGE is a diverse, high-quality dataset with samples resequenced from the 1000 Genomes Project. [12] If results from RPVG were promising, this dataset contains samples from all over the world, and the pangenome analysis would be run on the entire dataset.

20 Samples from the MAGE Dataset were run through MPMAP and RPVG and compared against, ASE data generated from standard ASE methods. The standard ASE data had been previously QCed and is a high-quality reference to see where ASE data produced by RPVG differs. These 20 samples run through RPVG, did not produce clean ASE data, and to save computational resources the rest of

the MAGE dataset was not run through the HPC cluster. One sample, HG00114 is shown here for the rest of the graphs, as all other samples, produced similar results, however, the figures for all samples are available in the supplemental materials.

Because RPVG produced counts on a transcript level, RPVG datasets were remapped back to gene level to be compared to the phASER output which was also on gene level.

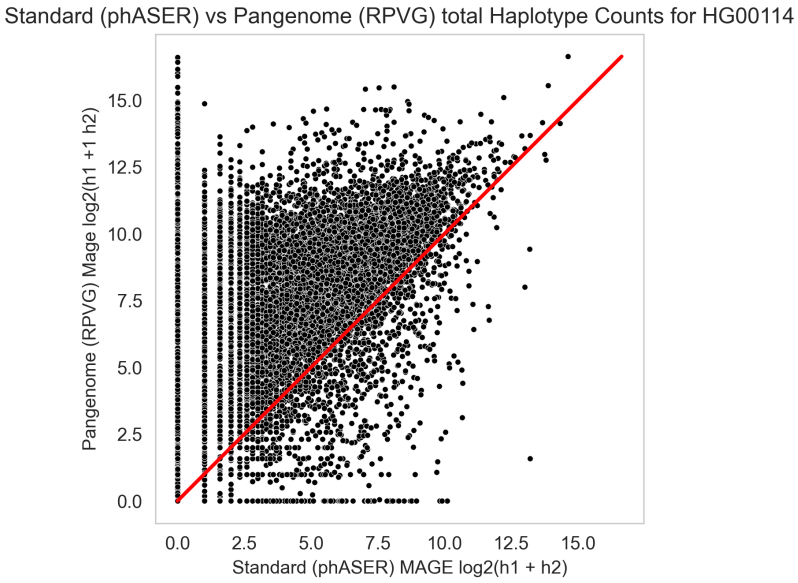


Figure 4: Pangenome (RPVG) vs Standard (phASER) Total counts for HG00114. A line with a slope of 1 shows where total counts for both methods are equal.

When running the RPVG pipeline on the MAGE dataset, RPVG and MPMAP produced significantly more counts than MAGE data generated with STAR and phASER. RPVG produced more counts for almost all genes compared to standard ASE methods. Banding patterns between 0 and 2 for the standard ASE data, are regions of the genome that are masked, and not counted, as they are repetative regions of the genome. These regions are difficult to align properly using traditional methods and are usually not counted.

A KDE plot of both datasets shows that the majority of the dispersion for both plots is around 0.5, which is expected as most genes should not be overexpressed. Figure 5, RPVG (Blue) had a much higher peak at the 0.5 mark, compared to phASER (RED), which was more dispersed around 0.5 when comparing expression levels of haplotype 1 divided by total haplotype expressed (haplotype 1 + haplotype 2). In the output files from RPVG, there were many points where both haplotypes had the same counts. This is surprising as RPVG already produced counts that contained fractional outputs. For the majority of genes to have an exact ratio of 0.5, compaired to phASER, which produces outputs as whole numbers and with a more natural dispersion around 0.5, is unlikely. It appears RPVG is estimating the haplotype ratios for these genes.

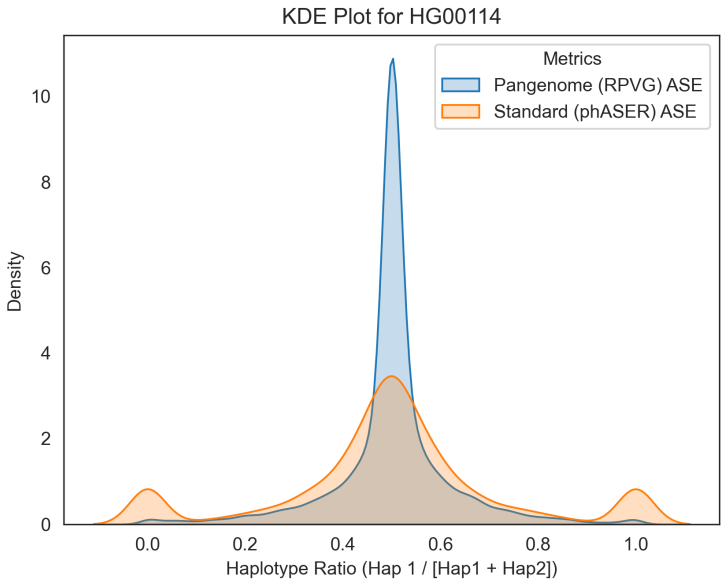


Figure 5: Kernel density estimate (KDE) to show density of haplotype expression ratio of two haplotypes. Haplotype 1 / Total Counts for Pangenome (RPVG, Blue) and Standard (phASER, Orange) total counts for HG00114.

4.3 ASE analysis of results

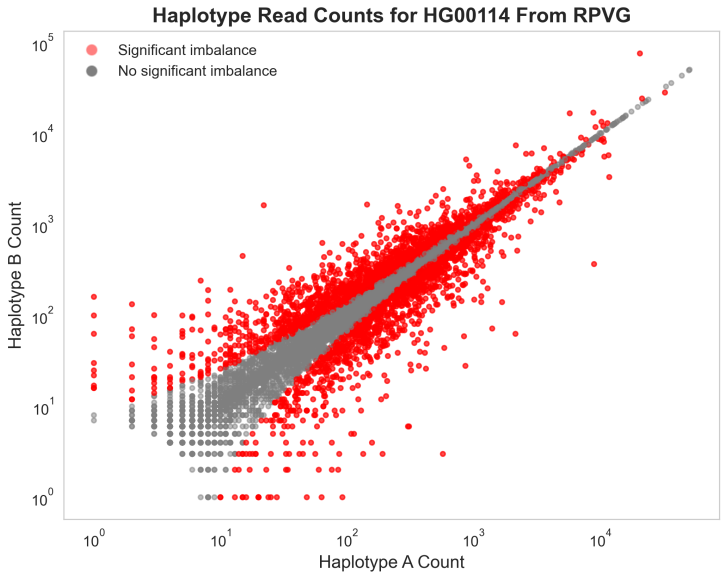


Figure 6: Comparision of haplotype expression with a binomial distribution for HG00114, produced with RPVG.

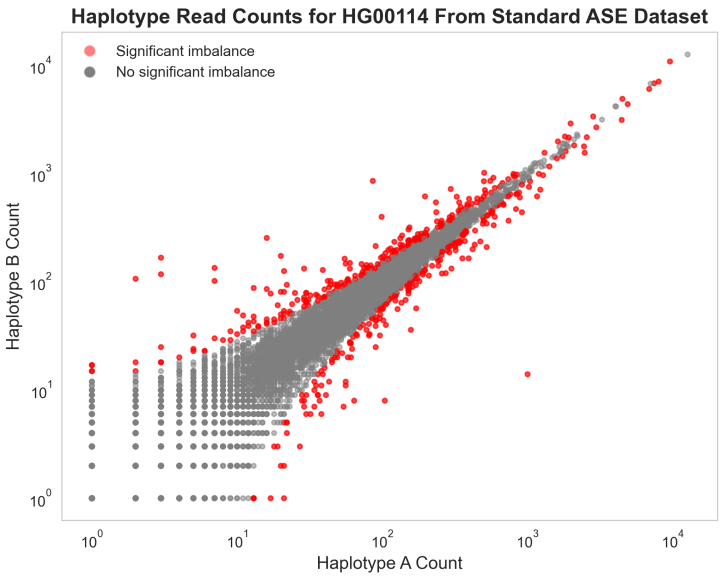


Figure 7: Comparision of haplotype expression with a binomial distribution for HG00114, produced with standard tools (phASER).

A binomial test was performed on the output of RPVG against the standard ASE analysis output from phASER, to measure the expression of each haplotype with points considered as overexpressed and outside of a 95 percent confidence interval highlighted in red. Figure 7 shows haplotype expression for data generated from standard methods and is an example of what standard ASE output generally looks like. Most genes are not expected to be overexpressed, and with most points falling along a linear trend, of a slope of 1, where expression levels are almost equal.

Figure 6 shows the output of RPVG, not only are more genes overexpressed, but the level of expression is much higher generally than the standard ASE data for the same sample. There is a lot more noise or dispersion in RPVG data, making it difficult to determine which genes are truly overexpressed.

5 Discussion

While RPVG is a promising tool, producing more counts does not necessarily translate to higher-quality data. Although RPVG produced significantly more counts than Kallisto or phASER in this analysis, as well as with other tools in the RPVG paper, further analysis told a different story.

ASE data produced by RPVG was significantly overexpressed, for all samples after a binomial test was performed. If the results of RPVG were taken at face value, this would suggest that all previous ASE analyses were done incorrectly and that more alleles than expected are expressed, and the level of expression for those genes would be much higher.

Furthermore, RPVG failed to reproduce QCed high-quality ASE data from a dataset generated using standard tools and methods. While it was expected that RPVG would not match the results of previous tools exactly and that there would be more counts for some genes, to produce such highly different results suggests that RPVG is not producing accurate reproducible results. RPVG also produced a large number of genes with exact same levels of expression, even though the output of RPVG contains fractional counts. This suggests that RPVG is trying to predict counts for a given transcript, but is unable to infer the other haplotype properly.

RPVG functions very differently than other tools, with similar functions as a genotype is not used for ASE analysis, and that RPVG predicts possible transcript paths based on how the reads map to the pangenome reference.

The additional counts produced by RPVG were likely due to a few factors. Using a graph reference with the pangenome alignment of RNA reads in MPMAP and later in the pantranscriptome, with RPVG likely mapped more reads that other tools were unable to, this is in part by design. The mapping algorithm however was likely too lenient, in the step where they allowed partial alignments to a transcript. This is also likely why their output for RPVG does not produce whole numbers for read counts, including fractional counts instead. Furthermore counting on a transcript level makes it difficult to compare to previous ASE analyses, which are commonly done on gene level. The conversion of the transcript to gene likely compounded the already high counts.

If the VG Team were to improve their tool, adding an option to produce counts on a gene level, would be beneficial for future studies. The VG team are very experienced in the field of Pangenomics, this is their first paper discussing ASE analysis and in the future, if they want to improve RPVG, working with experts in the field of ASE analysis is essential.

These results in this project were overlooked by the VG team, as they only analyzed one sample for ASE, and did not perform a proper ASE analysis. Instead, they compared their results to expected simulated results. It is unfortunate, that in its current state, RPVG, does not produce reproducible results as there is a lot of potential in the tool, and in using graph references for ASE analysis as a whole. Running MPMAP and RPVG is a much more straightforward process to produce ASE data, compared to standard methods using STAR, andASEReadCounter or phASER. With standard methods, input options need to be catered to each sample, and for haplotype analysis, results must be phased. RPVG, also may work better by aligning to repetitive regions of the human genome that are currently have to be masked with traditional ASE methods, as graph references may make alignments easier to these regions of the genome. RPVG is currently the only tool that can use Pangenome graphs as input for ASE analysis, and until RPVG is improved, or a new tool comes along, the issue of reference bias will need another solution.

6 Supplemental Materials

All code use the pipeline and used are figures are available on github. Additional figures for the rest of the MAGE dataset, are also available on the github:
`https://github.com/Cave42/Pangenome_ASE`

Sample ID	SRR ID
HG00265	SRR19762408
HG00260	SRR19762234
HG00254	SRR19762504
HG00250	SRR19762660
HG00244	SRR19762665
HG00243	SRR19762419
HG00237	SRR19762247
HG00233	SRR19762678
HG00160	SRR19762584
HG00151	SRR19762718
HG00148	SRR19762787
HG00146	SRR19762594
HG00142	SRR19762732
HG00141	SRR19762808
HG00132	SRR19762855
HG00130	SRR19762612
HG00127	SRR19762815
HG00122	SRR19762869
HG00121	SRR19762757
HG00114	SRR19762921

Table 2: MAGE Dataset Samples Selected, run and analysised for this project

References

[1] Stephane E. Castel, Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16(1), September 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0762-6. URL `http://dx.doi.org/10.1186/s13059-015-0762-6`.

[2] Stephane E. Castel, Pejman Mohammadi, Wendy K. Chung, Yufeng Shen, and Tuuli Lappalainen. Rare variant phasing and haplotypic expression from rna sequencing with phaser. *Nature Communications*, 7(1), September 2016. ISSN 2041-1723. doi: 10.1038/ncomms12817. URL `http://dx.doi.org/10.1038/ncomms12817`.

[3] Xian Chang, Jordan Eizenga, Adam M. Novak, Jouni Sirén, and Benedict Paten. Distance indexing and seed clustering in sequence graphs. December 2019. doi: 10.1101/2019.12.20.884924. URL `http://dx.doi.org/10.1101/2019.12.20.884924`.

[4] Nae-Chyun Chen, Brad Solomon, Taher Mun, Sheila Iyer, and Ben Langmead. Reference flow: reducing reference bias using multiple population genomes. *Genome Biology*, 22(1), January 2021. ISSN 1474-760X. doi: 10.1186/s13059-020-02229-3. URL `http://dx.doi.org/10.1186/s13059-020-02229-3`.

[5] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-

seq aligner. *Bioinformatics*, 29(1):15–21, October 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts635. URL http://dx.doi.org/10.1093/bioinformatics/bts635.

[6] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875â–879, October 2018. ISSN 1546-1696. doi: 10.1038/nbt.4227. URL http://dx.doi.org/10.1038/nbt.4227.

[7] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48(11):1443–1448, October 2016. ISSN 1546-1718. doi: 10.1038/ng.3679. URL http://dx.doi.org/10.1038/ng.3679.

[8] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20(9):1297–1303, July 2010. ISSN 1088-9051. doi: 10.1101/gr.107524.110. URL http://dx.doi.org/10.1101/gr.107524.110.

[9] Pejman Mohammadi, Stephane E. Castel, Beryl B. Cummings, Jonah Einson, Christina Sousa, Paul Hoffman, Sandra Donkervoort, Zhuoxun Jiang, Payam Mohassel, A. Reghan Foley, Heather E. Wheeler, Hae Kyung Im, Carsten G. Bonnemann, Daniel G. MacArthur, and Tuuli Lappalainen. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science*, 366(6463):351356, October 2019. ISSN 1095-9203. doi: 10.1126/science.aay0256. URL http://dx.doi.org/10.1126/science.aay0256.

[10] Pejman Mohammadi, Stephane E. Castel, Beryl B. Cummings, Jonah Einson, Christina Sousa, Paul Hoffman, Sandra Donkervoort, Payam Mohassel, Reghan Foley, Heather E. Wheeler, Hae Kyung Im, Carsten G. Bonnemann, Daniel G. MacArthur, and Tuuli Lappalainen. Quantifying genetic regulatory variation in human populations improves transcriptome analysis in rare disease patients. May 2019. doi: 10.1101/632794. URL http://dx.doi.org/10.1101/632794.

[11] Jonas A. Sibbesen, Jordan M. Eizenga, Adam M. Novak, Jouni Sirén, Xian Chang, Erik Garrison, and Benedict Paten. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. *Nature Methods*, 20(2):239–247, January 2023. ISSN 1548-7105. doi: 10.1038/s41592-022-01731-9. URL http://dx.doi.org/10.1038/s41592-022-01731-9.

[12] Dylan J. Taylor, Surya B. Chhetri, Michael G. Tassia, Arjun Biddanda, Stephanie M. Yan, Genevieve L. Wojcik, Alexis Battle, and Rajiv C. McCoy. Sources of gene expression variation in a globally diverse human cohort. *Nature*, 632(8023):122–130, July 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07708-2. URL http://dx.doi.org/10.1038/s41586-024-07708-2.