**Epigenetic Age in Age-Related Macular Degeneration WIP**

**[slide 2]**

So I’ll start with a brief background of what AMD or Age-Related Macular Degeneration is. AMD is a common degenerative disorder that affects the central retina of those over 50 years old and is the leading cause of blindness in several populations. The disease is characterised by the accumulation of fatty deposits called drusens in the central retina which results in vision impairment as shown in this figure here. It primarily exists in three forms dependening on the age of onset - early, intermediate and late AMD.

**[slide 3]**

The different forms of AMD are characterised by the size and accumulation of drusen in the retina. In the figure above, you can see the gradual increase in drusen size with disease severity.

Early AMD is characterised by the presence of small or intermediate drusens (small = less than 63 microns; intermediate = 63 microns - 125 microns) in the retina. Note that small drusens are frequently observed in patients over 50 years old and can represent an epiphemenon of aging.

Intermediate AMD is characterised by intermediate or large drusens (63 microns - 125 microns) and...

Late AMD is characterised by the widespread accumulation of large drusens (more than 125 microns) in the retina.

Importantly, late AMD is subdivided into two subtypes: Geographic atrophy (commonly known as dry AMD) and choroidal neovascularisation (commonly known as wet AMD). Dry AMD gets its name as it is caused by the accumulation of hardened drusens (which are dry fatty deposits), while wet AMD is caused by the leakage from blood vessels that grow under the retina. Because of the nature of choroidal neovascularisation, it is still possible to treat the disease, while there is currently no known treatment for geographic atrophy.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** Late AMD (Wet) - Present in one eye 1.2% of the time; (Dry) - Present in one eye 0.6% of the time.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 4]**

Much of this research has focused on the factors contributing to AMD pathology which include several environmental and genetic risks. In a 2005 study investigating 840 male twin pairs from the US (440 monozygotic twins - share complete genetic makeup, and 400 dizygotic twins - share half of their genes), it was determined that the heritability of AMD is 46% and 71% for early and late AMD respectively. Since 46% - 71% of AMD variation may be explained by genetic factors, they have been extensively explored in the context of GWAS studies. Conversely, the heritability of AMD also highlights the relative importance of environmental risk factors such as smoking and diet in the early stages of the disease. This provides an avenue for disease exploration in an epigenetic context as it has been suggested that tissue-specific DNA methylation differences may be the key in understanding the interplay between genetic and environmental risk factors in AMD.

Ref: Seddon et al. (2005) The US Twin Study of Age-Related Macular Degeneration.

Twins from World War II Veteran Twin Registry.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** On monozygotic and dizygotic twins - a trait influenced entirely by additive genetic factors should correlate perfectly in MZ twins and approximately 0.5 in DZ twins. A trait influenced entirely by shared environmental factors should correlate perfectly in both MZ and DZ twins. ACE (additive genetic, common environment and unique environment model) predicts DZ correlation ranging from 0.5 of MZ correlation (i.e. purely additive and unique environment) to being equal to MZ (i.e. purely common and unique environment).

Unique environmental factors accounting for 19% - 37% of the variation in AMD grade. No evidence for shared environmental factors (however, twins most likely had not been cohabiting for over 50 years, so it’s not surprising that shared environmental effects are not detected).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 5]**

Risk of AMD increases with age, with 3-fold increases of AMD incidence amongst individuals over 75 compared to individuals between 65 - 74.

Smoking is the most consistent risk factor in AMD where relative risk has been reported to decrease by 16% (2.4 -> 2.0) with smoking cessation.

There is also a global effect of diet in the incidence of AMD. Low dietary antioxidants and a high-fat diet can lead to increased risk of AMD by affecting cardiovascular pathways.

Early AMD is commonly observed among Black and Hispanic populations, while late AMD is much more common in White populations.

Diagnoses of late AMD is slightly higher in women compared to men. However, it is important to note that this observed difference is in neovascularised AMD in women of European ancestry specifically. Comparing overall AMD in men and women does not show any substantial difference.

In this project, I’ll be investigating the role of age in AMD in the context of epigenetics. Specifically, I will characterise the incidence and progression of AMD using epigenetic clocks.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:**

**Why smoking?** Smoking adversely affects blood flow, decreases levels of high-density lipoprotein, increases platelet aggregability and fibrinogen, increases oxidative stress and lipid peroxidation, reduces plasma levels of antioxidants, and raises levels of inflammation and inflammatory cytokines.

**Why high-fat diet and antioxidants?** The deposit of oxidised compounds in healthy tissue could result in cell death because they are indigestible by cellular enzymes in the eye. This could lead to impaired function of the RPE and eventually to degeneration involving the macula. High-fat diet likely affects AMD through the HDL cholesterol and systemic inflammatory pathways (See above).

**Why ethnic differences in AMD?** Inherited genetic factors determined by genetic ancestry play a substantial part in age-related macular degeneration. Also, most studies are localised to specific geographical locations, so it is suggested that there could be some sampling bias.

**Why more late AMD in women?** The higher prevalence of **late AMD** in women is due to potential cerebrovascular and/or hormonal/menopausal mechanisms: Neovascularised AMD seems to share some similarities with stroke. Women experience their first strokes later in life and have higher incidence of stroke at an age when neovascularised AMD commonly occurs (an age at which male cardiovascular-related mortality may have already occured = in short, because women live longer than men). [But important to stress that no evidence of any substantiated difference between men and women in overall AMD] (See below).

**NOTE:** A meta-analysis of 25 published studies (over 57000 subjects) identified no evidence of interaction between age and gender (i.e. prevalence of AMD increases exponentially by age, which did not differ by gender).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 6]**

This leads to my main question for this project, which is: Can the increased risk of AMD with age be partially characterised with epigenetic clocks? I expect that samples from AMD patients will have higher age acceleration as compared to controls.

**[slide 7]**

To help me answer my question, I’ll briefly walk you through the publicly available dataset that I’m using and the pre-processing steps that I have undertaken to clean it up for analysis.

**[slide 8]**

The dataset I will be using contains 44 samples with ages ranging from 50 to 89 years old. The sex and disease state of the samples are quite evenly split. All the samples from the dataset comes from the retinal pigment epithelium with no health history or cause of death provided, and the donors are most likely based in the UK as all tissues were from the Manchester Eye Bank in the UK.

To prepare this data I used our Kobor lab pipeline, importing the raw idats and metadata from ArrayExpress (which is a European-based data repository for microarray and sequencing datasets) into R and performed some preliminary exploration of the available metadata.

**[slide 9]**

Preliminary PCA on the raw data shows that sex is the major source of variation in this dataset. It is interesting to see that age does not show up in the top PCs and disease state does not show up at all, so further exploration is needed to understand why that is.

**[slide 10]**

Because it is a little concerning that age is not one of the main sources of variation as we typically see, I looked at a correlation matrix between all the variables that are available in the metadata. From this plot, you can see that age is correlated with chip and disease state is correlated with row, this confounding is concerning as age is the main variable that I intend to investigate.

**[slide 11]**

Next I will quickly walk you through the checks that I have performed to filter out any bad samples. I performed an extensive check on these biological and technical metrics to ensure that the samples are up to par. Fortunately, all samples passed the checks so I did not need to discard any sample from this dataset, as the sample size is already quite low.

**[slide 12]**

As Horvath recommends using pre-normalized data, I only performed noob-normalisation to correct for background fluorescence and dye bias, as was suggested by David. The resulting distributions overall and between probe types are more comparable to one another after this step.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** Noob normalisation uses out-of-band probes on the array to estimate non-specific signals from the out-of-band intensities (which are wavelengths that are in the opposite colour channel to their design). These signals are then normalised in average intensity in the red and green colour channels.

Potential argument for a different choice of normalisation - While it is true that the choice of normalisation can introduce some variation by influencing the mean values of DNAm age, the predicted ages demonstrated in Lisa’s paper are still within the error margin of the clocks.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 13]**

After normalisation, I uploaded my dataset onto Horvath’s online calculator to produce readings for various epigenetic clocks. As the source tissue of this data is retinal pigment epithelium, I chose Horvath’s because it is a pan-tissue clock and so theoretically should be able to predict age in these samples. Since the tissue is so unique and not included in any of the training data for any currently available clocks, I decided to try my tissue with most of them to see how well they perform with my tissue. Overall, I explored 7 epigenetic clocks in total.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** Epigenetic clocks use the methylation profile of age-associated CpG sites to approximate the biological ages of the samples using DNAm differences. While the biological age predicted is usually quite correlated to chronological age, differences between predicted and actual age has been associated with various health outcomes – this measure is known as epigenetic age acceleration.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 14] (change bottom line on slide to “Age acceleration\* is higher for AMD samples as compared to normal samples.”)**

I expect the predicted biological ages of AMD samples to be higher than normal samples using the epigenetic clocks. I also expect the age acceleration of AMD samples to be higher as compared to normal samples. In this presentation, you will see two prominent terms that are recommended by Horvath to be a more useful characterisation of epigenetic age, which are age acceleration differences and age acceleration residuals.

Age acceleration differences is the difference between predicted biological age and chronological age, and

Age acceleration residual are the residuals resulting from regressing predicted biological age with chronological age. Positive residual values indicate faster aging in relation to chronological age, and negative residual values indicate slower aging.

**[slide 15]**

Before I present you with the epigenetic clock results, I’ll share some characteristics of the clocks with you. All the clocks differ by the age-associated CpG sites that they are built on. They also have vastly different sample sizes and tissue types for their training datasets. The most well-known epigenetic clock is Horvath’s clock, which is said to be able to predict the biological age of any tissue as it was built using 24 different sources. There are some tissue-specific clocks that are more accurate in their prediction within tissue types that they are trained on such as Zhang’s clock and the Cortical clock. There is also an increasing development of so-called second-generation clocks that are trained on not only chronological age, but also age-related and disease phenotypes such as mortality. For example, the PhenoAge clock incorporates nine age-related biochemical measures, which allows the clock to capture age-associated phenotypes.

The clocks that I have highlighted contain tissue types in their dataset which the clocks that I hypothesise to be more similar to my tissue type in their trained tissues, and so I expect their prediction accuracy to be better than the other clocks.

Why did I choose Horvath’s clock: Pan-tissue.

Why did I chose Skin&Blood clock: Said to be able to accurately predict neurons, glia, brain.

Why did I choose Cortical clock: Accurately predict different brain cells.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:**

**Why was Skin&Blood clock built:** To overcome the sub-optimal prediction of Horvath’s clock for fibroblast tissues and to address a need for a fibroblast predictor that is widely used in ex vivo intervention studies.

**Why was Cortical clock built:** Existing clocks work sub-optimally in the human cortex, systematically underestimating age in elderly individuals.

**Why was Zhang’s clock built:** Argues that predictors based on small sample sizes are prone to confounding by cellular compositions (said that Horvath & Hannum’s clocks were enriched in CpGs that show DNAm heterogeneity across cell types), so wanted to build a predictor that best capture underlying age-associated sites can robustly estimate epigenetic age.

**Why was Hannum’s clock built:** Hannum predates Horvath’s clock, initially to characterise aging in blood.

**Why was Weidner’s clock built:** Hannum predates Horvath’s clock, wanted to see if they can find a cost-effective method of characterising biological age using a little CpG sites as possible - using bisulfite sequencing. Horvath thinks that their age prediction is rubbish - high error rate (up to 13 years).

**Why was PhenoAge clock built:** Wanted to build a clock that can predict various aging outcomes - all-cause mortality, cancers, healthspan, physical function and Alzheimer’s.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 16]**

After I applied the clocks to my dataset, I saw a systematic underestimation of the predicted biological ages of my samples across all clocks.

Among all epigenetic clocks tested, Horvath’s clock was able to produce predicted biological ages that are as close to the actual chronological ages as possible. However, as you can see from the regression plot next to you, it is still resulting in large underestimation. The y-axis in this plot shows the predicted biological ages and the x-axis shows their actual chronological ages. As you can see, some samples such as this [points] shows a 20 year predicted age, which is way lower than their 50 year chronological age.

**[slide 17]**

Next, I coloured the plot points by sex and change their shapes to reflect their disease state to show if any of these phenotypes are driving these differences in ages. Except for this normal sample here, it seems like perhaps the normal samples are closer to the regression line as compared to the AMD samples. But when I checked the statistical significance of this effect, I see that this is not significant.

**[slide 18]**

So what happened there? I started to look in literature to see if there are other tissues that are systematically underestimated as well. Incidentally, the group that built the Cortical clock saw similar underestimations in their datasets of brain tissues, specifically amongst those of older ages. They saw that the clock started to underestimate the brain tissues amongst individuals over 60 years old as seen in this Bland-Altman plot. The expectation from a Bland-Altman plot is that when the means of the predicted age and actual age against the differences between them, we will see a strong pattern close to 0 if the predictions are accurate. Any deviations from 0 show inaccurate predictions. I plotted a similar Bland-Altman plot, coloured by disease status, to see the predictive performance of my tissue as well and saw similar deviations. This is interesting because Horvath’s clock is trained with some of the brain tissues that they are using, so the common denominator here seems to be the underestimation of older ages. They proposed two things that could possibly address this pattern that they are seeing: (1) Horvath’s clock is built without an equal representation of the 0 - 100 years age range across all tissues. This means that the predictive ability of the clock is limited by the range of age of the tissues that were used to train it. (2) They suggest the possibility of non-linear aging later in life. They propose that as we age to later in life, the age-associated CpG sites that are characterised by the clocks slowly reach full methylation or complete demethylation, and thus do not represent the methylation patterns captured by the clocks.

The Bland-Altman plot for blood is equally interesting as we see all the points concentrated around 0. This brings up an interesting idea of looking at tissue-specific effects, as we see here that the aging of blood is more accurately predicted than the other tissues.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Only if asked:** Explanation of Bland-Altman plot - The Y-axis of a Bland-Altman plot shows the difference between two quantitative measurements (A-B) and the X-axis shows the mean of the two quantitative measurements ((A+B)/2). In other words, the difference between the measurements is plotted against the mean of the measurements. B&A recommends that 95% of all the points should lie within +/- 2 standard deviation of the mean difference. (Giavarina 2015).

Giavarina, Davide (2015) Understanding Bland Altman Analysis. Biochem Med (Zagreb).

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**[slide 20]**

For that I revisit the regression plot and see that even though the ages are underestimated, the predicted ages are significantly correlated with the chronological ages. As the correlation is significant and high, using acceleration measures may still be notable and worth investigating.

**[slide 21]**

I hypothesise that there may be some tissue specific markers that are shared between my tissue and the tissues that are used to train Horvath’s clock. Retinal pigment epithelium develops from the same germ line as other tissues used in Horvath’s clock such as skin and neuronal tissues. In this diagram below, it describes the development of RPE from an optic vesicle into a more differentiated RPE layer. These authors suggest that this layer was derived from an ancestral ectodermal cell containing light-sensitive pigments into the anterior neuroectoderm that eventually derives into the optic neuroepithelium and then the RPE. Similarly, this neuroectoderm has also been described as the progenitor of neural tissues such as neurons and glia (Hartenstein 2015). Because of this shared lineage, I suspect that these tissues share some similar tissue-specific markers initially, but are eventually lost due to differentiation into the highly specialised tissue in RPE.

Hartenstein & Stollewerk (2015) The Evolution of Early Neurogenesis.

**[slide 23 & 24]**

If you recall, a positive age acceleration value denotes an accelerated age with respect to the sample’s chronological age and vice versa. When I plotted the age acceleration residual, I saw that the age acceleration residual was slightly higher for AMD samples than the controls but the difference was not significant. I then stratified the samples by sex and saw that the difference was driven by the male AMD samples. To look at the effect of the male sample, I plotted the interaction to see if there are any interaction effects between the two variables, and saw that there is an interaction, but the interaction is not statistically significant.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** An interaction plot displays the fitted values of the dependent variable on the y-axis while the x-axis shows the values of the first independent variable in the interaction term. The lines represent the values of the second independent variable. In an interaction plot, parallel lines indicate that there is no interaction effect while different slopes suggest that one might be present. When the lines cross each other, the interaction effect is strong. Also, when the line is parallel to the x-axis, there is no main effect present. The response mean is the same across all factor levels. When the line is not horizontal, there is a main effect present. The response mean is not the same for all factor levels. The steeper the slope of the line, the greater the magnitude of the main effect. HOWEVER, non-parallel lines do not necessarily represent an actual effect. It can also represent random sample error. A statistical test can help to understand if the effect is significant.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 25]**

Picking up on the difference in aging rate in RPE as compared to other tissues in the Bland-Altman plot, I thought it might be interesting to look for any age-associated sites in RPE. Aging in RPE has been explored in other context, but not as extensively from an epigenetic perspective.

**[slide 26 & 28]**

To do that, I attempt an EWAS to see if there are any age-associated sites that are attributed to RPE. In the literature, ELOVL2 has been identified as an age-associated site in the retina of mice. I hope to find similar sites in my exploratory study as well.

**[slide 28]**

Picking up from sample quality control, I performed funnorm on my data. Briefly, I chose funnorm for my EWAS because funnorm is useful for studies comparing conditions with known differences, like a case-control study. Because it is a between-array normalisation method, it will adjust my Type I distribution to be more similar to my Type II distribution, as you can see here.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** Funnorm is a between-sample (functional) normalization method that attempts to remove unwanted variation by adjusting for covariates estimated from a control probe matrix. Briefly, 42 summary measures are estimated from the combined 848 control probes and type I ‘out-of-band’ intensities, with the first *m* = 2 principle components of the summarized measures chosen as covariates for intensity adjustment. Adjustment is performed separately in methylated and unmethylated intensities, and in type I and II probes. For probes mapped to X and Y chromosomes, males and females are processed separately, with ordinary quantile normalization used for probes on the Y chromosome because of the small number of probes (*N* = 416). By default the functional normalization is applied after Noob in the current version of minfi package.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 29]**

Then I proceeded with our standard probe filtering methods, where XY, SNP, polymorphic, cross reactive and probes that did not pass detection p-value and bead count are removed.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 30]**

After that, I performed an SVA to account for any unknown biological variation in my dataset. I did not account for technical variation because the batch effects are known. A large issue with this study is the batch and variable confound I mentioned earlier. I can either adjust for batch and lose true hits or keep batch and increase my false positives, neither choice is desirable and so I simply chose to opt out of performing ComBat altogether, the approach used by the original authors. However, any results must then be interpreted with caution. The SVA algorithm identified one significant SV, which was added to my PC to see where it shows up, and it showed up as the first PC.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** SVA is a supervised method used to identify potentially large number of environmental and biological variations (and technical variations like batch) that are unmeasured and may have a large impact on measurement. It identifies and estimates the unmeasured variation as a surrogate variable while protecting the variables of interest.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 31]**

Now that I know my SV is contributing to a large source of variation, I need to identify what it is. I suspect that it may be attributed to the underlying cell type composition or even genetic differences, but unfortunately there is currently no available method to predict cell type for RPE tissues and the authors didn’t report ethnicity. My attempt at addressing which the SV may be associated to resulted in my using EpiDISH to predict cell type proportions as it is used to epithelial cell type differences, and because my tissue is a type of highly differentiated epithelial tissue, I hope that it is able to identify any cell type differences in my tissue. Using the EpiDISH algorithm, I saw that it predicted a high proportion of epithelial cells and fibroblasts, while the blood cells remain low, which is good.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Only if asked:** EpiDISH infers the underlying cell type proportions using robust partial correlations of the CpGs that are mapped to the reference database that was identified using a supervised selection procedure that looks for DHS (DNase Hypersensitive sites) and differentially methylated CpGs for each cell type.

DHS sites are lymphocyte specific, indicating that these sites can be used to identify gene regulatory elements that control cell type specificity.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**[slide 32]**

I see that predicted cell types show up as the top PC with the SV, so I suspect that they may be correlated.

**[slide 33]**

So, using a correlation matrix, I see that my SV and specifically, the epithelial cell are correlated. This allows me to infer that SV can be interpreted as a potential surrogate for intraindividual epithelial cell proportion differences in my data.

**[slide 34]**

After that, I look at the overall correlation between all my variables, and I still see the correlation between age and chip and disease state and row in my dataset.

**[slide 35]**

Because I’m interested in age as my main variable, I plot the ages across all chips and see that the ages are not evenly distributed across all chips.

**[slide 36]**

This makes me think back about the preprocessing steps described in the paper where I took this dataset from, because I realised that they had skipped ComBat, so I suspect that they decided to skip the step to not remove any disease state and age-associated effects in their EWAS.

**[slide 38 & 39]**

After understanding the potential batch effect as a confounder in my dataset, I performed an EWAS using the linear model [shown on my slide], explicitly accounting for my cell-type differences through my SV and the chip effect (which is confounded with age). I identified two significant increased methylated CpG sites (one with FDR-adjusted p-value less than 0.1 and the other with FDR-adjusted p-value less than 0.2) with delta beta more than 5%. One of the sites is mapped to LOC100132215 gene, which was previously identified and validated as significantly age-associated in the whole blood of 400 individuals aged between 50 to 75 years old while the other site is not located in a known gene region. Unfortunately, I am not able to identify the non-tissue specific age-associated ELOVL2 as one of the sites. However, given the limitation of the study design and low sample size of this dataset, it is interesting that the site that is identified as age-associated here is also identified in another tissue within the same age-range in a much better powered study, opening up the possibility of this site being identified as another potential non-tissue specific age-associated CpG.

**[slide 41]**

With that..., acknowledgements and thank yous.