Clara_ALD_Report

David Lin

4/4/2018

This Report summarises the pipeline and approach used to identify differentially methylated CpGs/regions in the x-ALD pilot project.

Study design

The Illumina MethylationEPIC arrays were performed on 12 DNA samples, extracted from purified lymphocyte samples. These samples belong to 12 different individuals, consisting of 6 discordant sibpairs in terms of their disease progressions (see **Table 1** for metadata). The 12 bisulfite-converted samples were randomly distributed across 2 chips and performed all in 1 batch (see **Table 2** for array layout). Following import into GenomeStudio, signals were color-corrected and background subtracted, and the resulting data were carried forward to the Kobor lab epigenetic pipeline as described below.

Table 1 Metadata for the x-ALD pilot project cohort

Sample_ID	cALD	Sib_Pair	Sex	Age.Now	Age	Ethnicity	Variant	ABCD1.mutation
ALD10	Yes	1	M	28	28	Caucasian	c.1390C>T	p.Arg464*
ALD11	No	1	M	29	28	Caucasian	c.1390C>T	p.Arg464*
ALD26	Yes	2	M	31	30	Caucasian	c.1899delC	p.Ser633Argfs*3
ALD27	No	2	M	31	30	Caucasian	c.1899delC	p.Ser633Argfs*3
ALD36	Yes	3	M	36	36	Caucasian	c.1992-2a>g	p.Lys665fs*?
ALD65	No	3	M	39	38	Caucasian	c.1992-2a>g	p.Lys665fs*?
ALD42	Yes	4	M	6	6	Caucasian	c.659T>C	p.Leu220Pro
ALD41	No	4	M	9	8	Caucasian	c.659T>C	p.Leu220Pro
ALD49	Yes	5	M	18	16	Caucasian	c.1866-2a>t	p.Pro623fs*?
ALD48	No	5	M	20	18	Caucasian	c.1866-2a>t	p.Pro623fs*?
ALD58	Yes	6	M	29	27	Caucasian	c.892G>A	p.Gly298Ser
ALD59	No	6	M	26	25	Caucasian	c.892G>A	p.Gly298Ser

Table 2 Array layout: Sample positions

Sample_ID	Sample_Plate	Sentrix_ID	Sentrix_Position	cALD	Sib_Pair
ALD36	WG6761599	201496850198	R01C01	Yes	3
ALD59	WG6761599	201496850198	R02C01	No	6
ALD58	WG6761599	201496850198	R03C01	Yes	6
ALD11	WG6761599	201496850198	R04C01	No	1
ALD26	WG6761599	201496850198	R05C01	Yes	2
ALD27	WG6761599	201496850198	R06C01	No	2
ALD10	WG6761599	201496850198	R07C01	Yes	1
ALD65	WG6761599	201496850198	R08C01	No	3
ALD48	WG6761599	201496860156	R01C01	No	5
ALD42	WG6761599	201496860156	R02C01	Yes	4
ALD41	WG6761599	201496860156	R03C01	No	4
ALD49	WG6761599	201496860156	R04C01	Yes	5

1. Sample Quality Check and Preprocessing

A. Sample Quality Assessment

Beta values, which represent percentage of methylation assigned at each of the 867,926 CpGs on the EPIC array, are examined prior to sample preprocessing. Typically a bimodal distribution is expected - which is what we see here in **Figure 1**

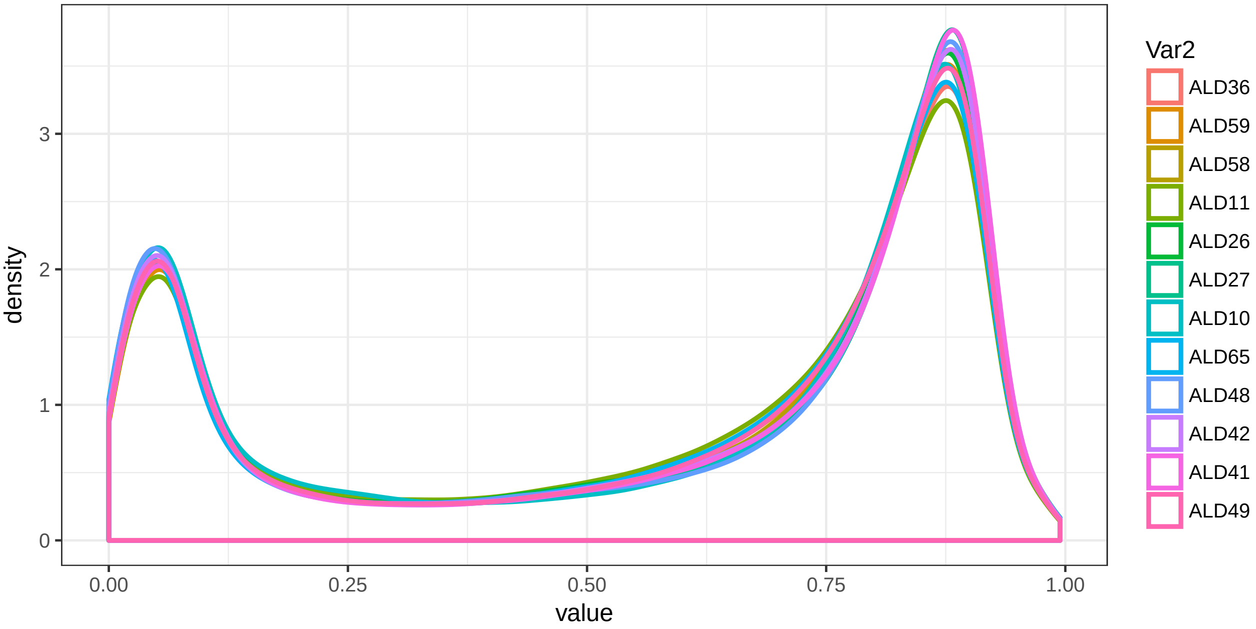

Figure 1 Raw sample beta distributions of the x-ALD pilot samples.

We also examined the overall sample detection p-values. A High detection p value indicates a failed sample; however in this instance all samples showed very low detection p values suggesting a successful run and therefore are kept in for the analyses that follow.

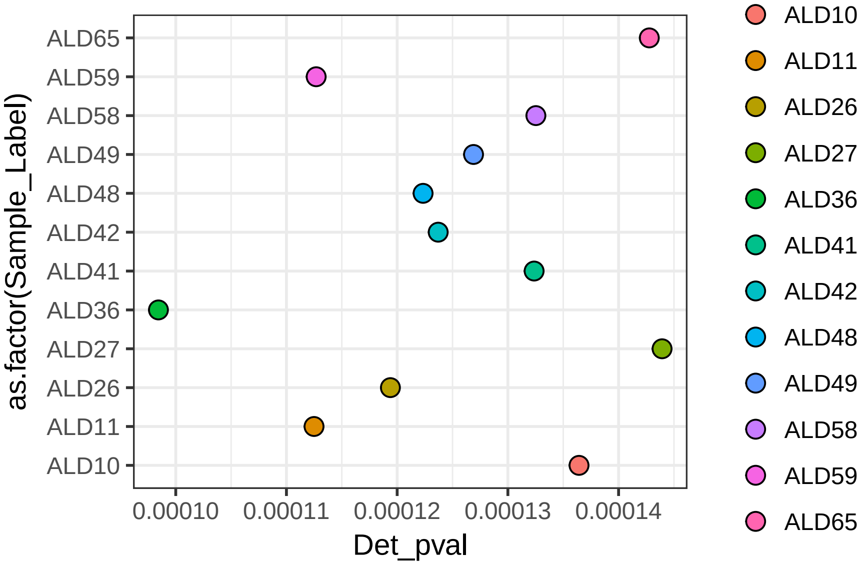

Figure 2 Sample Detection p-values for the array.

The EPIC array contains 59 SNP probes that allows us to perform a clustering analysis, confirming the identity of each sample:

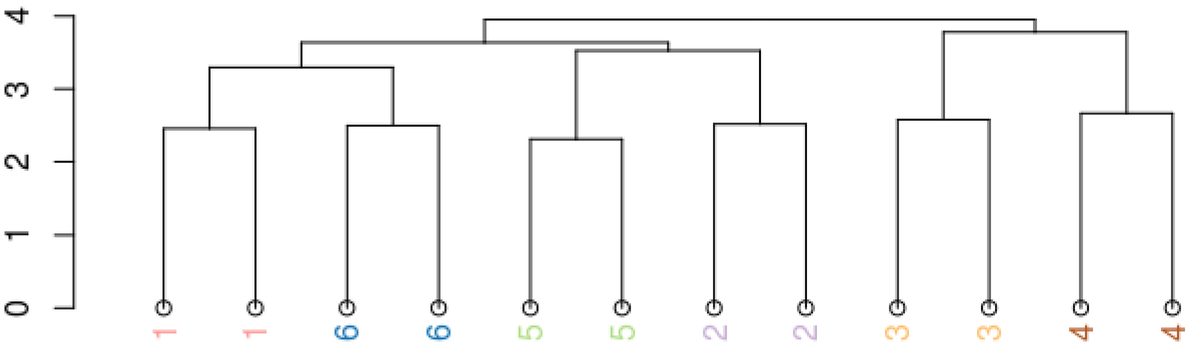

Figure 3 SNP Clustering of the samples. Each sample represents the age of the participant, basically fitting that all 4 treatments to a unique cell line (derived from different patient) cluster together.

B. Sample Preprocessing

Probe filtering

Using an established pipeline in the Kobor lab, I performed the following preprocessing steps to remove probes that are either poorly designed (ie. cross-reactive) or showed poor detection p-values across samples. **Figure 4** is a probe attrition plot that summarises the steps taken and the number of probes retained/removed through each stage.

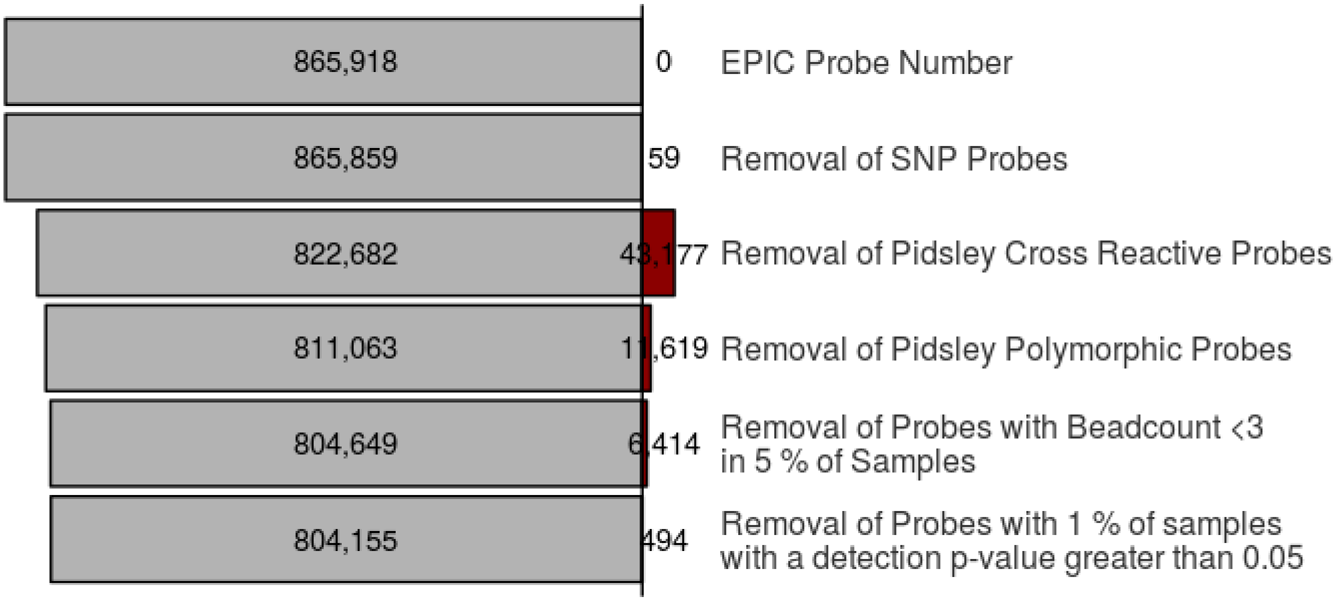

Figure 4 Sample attrition plot. The remaining 804,155 probes are carried forward to normalization.

Normalization

Following probe filtering, samples were normalized using the *DASEN* method in the WateRmelon package. DASEN works by normalizing the Type I and Type II probes separately. Normalization should bring sample distributions closer together, making the comparisons more fitting for analysis. *Note by David: Normally I would have used BMIQ, but the M-value transformed values look really drastic compare to what is normally seen, so I opted for DASEN*

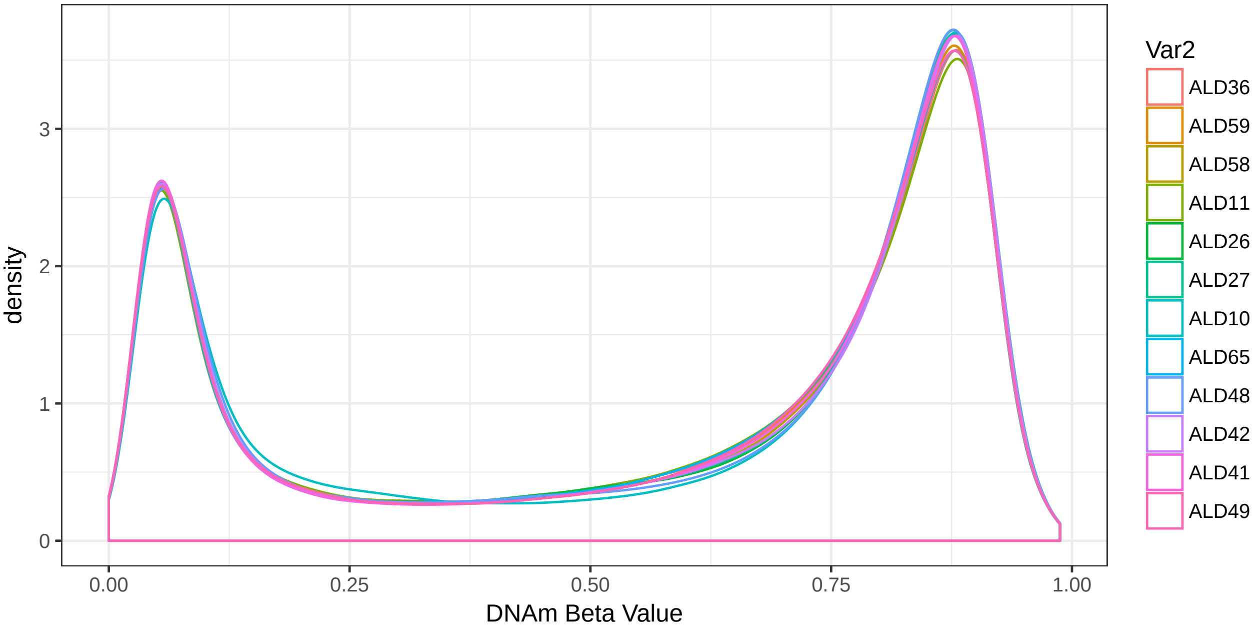

Figure 5 Sample beta distributions following normalization by the *dasen* method.

Examination of Sample Variations (PCA)

Next, we performed principal component analysis (PCA) to learn which variable contributes to the variation of DNA methylation that we observe in our data set. Using the provided metadata, we pulled out: cALD status, Sib pair, chip ID (“Sentrix ID”) and chip position (“Sentrix Position”), and age as the relevant metadata. We also included Sib-Pair as a numerical class rather than a factor (“Sib_Number”). I performed the PCA with these information on the normalized beta values. Note that our in-house PCA script also performes an association test to test for strong associations between any given variable in each PC.

As shown in **Figure 6**, Patient Age and Sib-ship (ie identity) are found in PCs2 & 3, which is expected as these factors are known to be associated with variable DNA methylation.

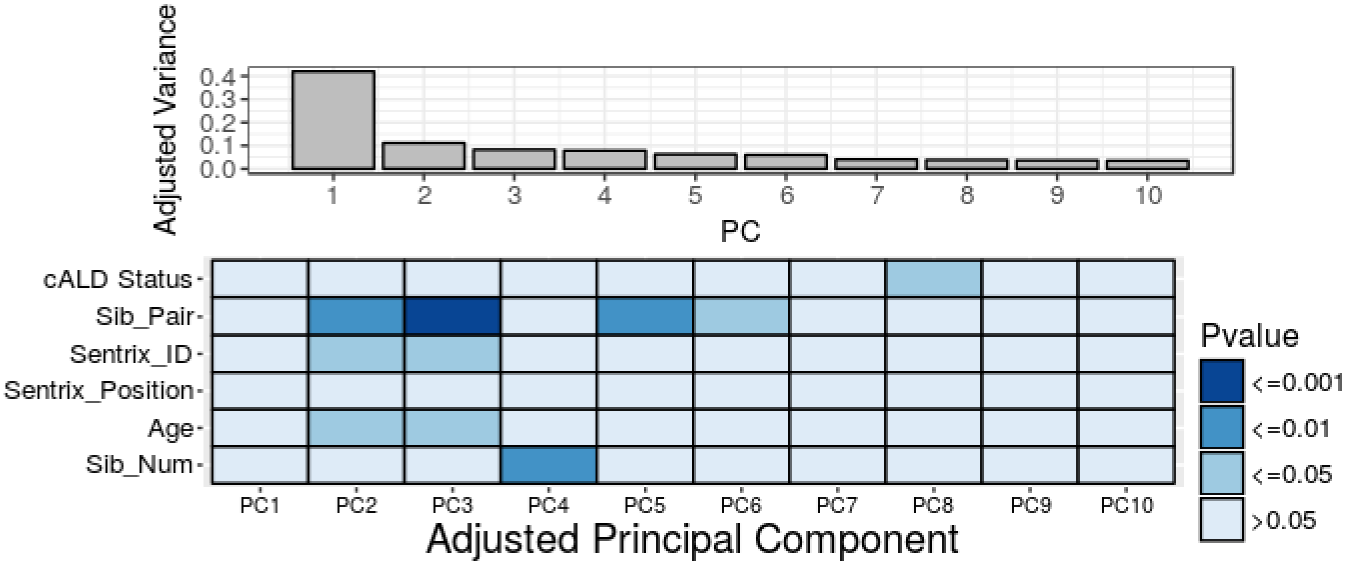

Figure 6 Principal Component Analysis (PCA) on the pre-processed and normalized x-ALD data set.

We note that Sentrix ID and Position show up in the different PCs, suggesting that these contribute to a batch effect. We were able to correct for Sentrix ID using ComBat. This is shown in **Figure 7**. Note unfortunately we did not have enough samples to meet the requirement to ComBat out Sentrix Position, which now shows up in PCs 2 and 4 following chip correction.

We can see that Sibship is still a major contribution to the DNA methylation variation observed. Whereas cALD condition is now in PC7.

Note ComBat fails when probes show 0 variance so these probes (244 of them) are removed, making the current probe count **803,911**.

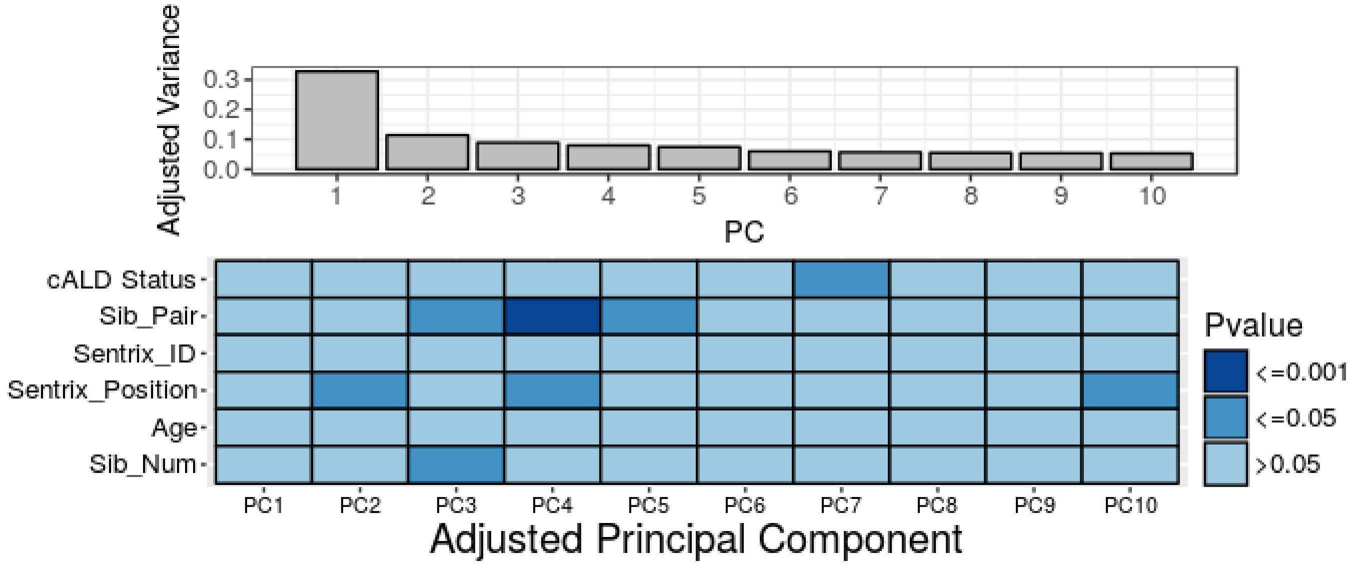

Figure 7 Principal Component Analysis (PCA) on the pre-processed and normalized data set, after ComBat to remove chip (Sentrix ID) effects.

Blood Cell Type Deconvolution

Purified lymphocyte samples, such as ones collected in the x-ALD pilot study, consist of several classes of agranulocytes, including T cells (CD4+ Helpers, CD8+ cytotoxic), B cells, and natural killer cells. Agranulocytes also includes monocytes. These lineages differ significantly in their DNA methylation profiles and the mixture of them, at unknown concentrations, may confound DNA methylation analyses for our main question (phenotypic differences), especially since differences in immunity may cause these cell proportions to fluctuate.

To correct for differences in these blood cell types, we performed bioinformatics-based blood cell prediction based on the DNAm profile of these samples,to estimate cell type in each sample, using a reference range method (see *Houseman et al., 2012*; **Figure 8, Left Panel**). A deconvolution procedure then follows to correct for these cell type differences (**Figure 8, Right Panel**).

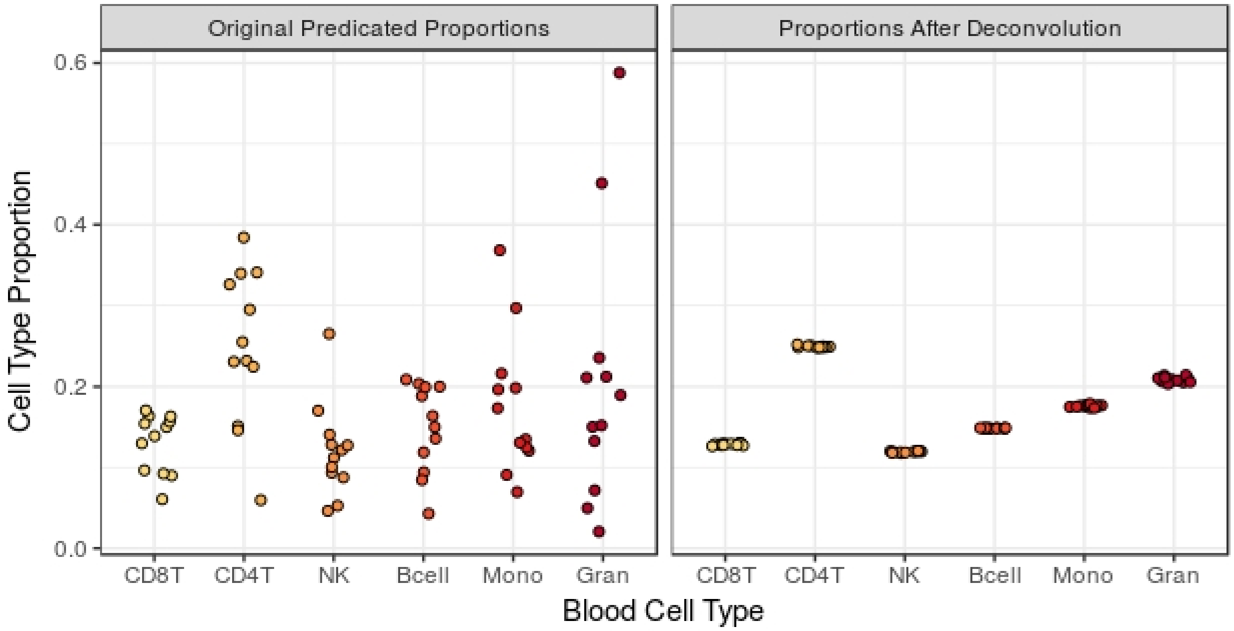

Figure 8 Estimated Cell Type proportions of each sample based on DNAm before (*left panel*) and after (*right panel*) cell type deconvolution.

Surprisingly, from **Figure 8**, we see a large proportion of granulocytes. However, the cell type proportions of the rest of the cell types were close to expected (ie. CD4+ being double the amount of CD8+ cells).

Figure 9 displays a summary of blood cell proportion for each of the individuals.

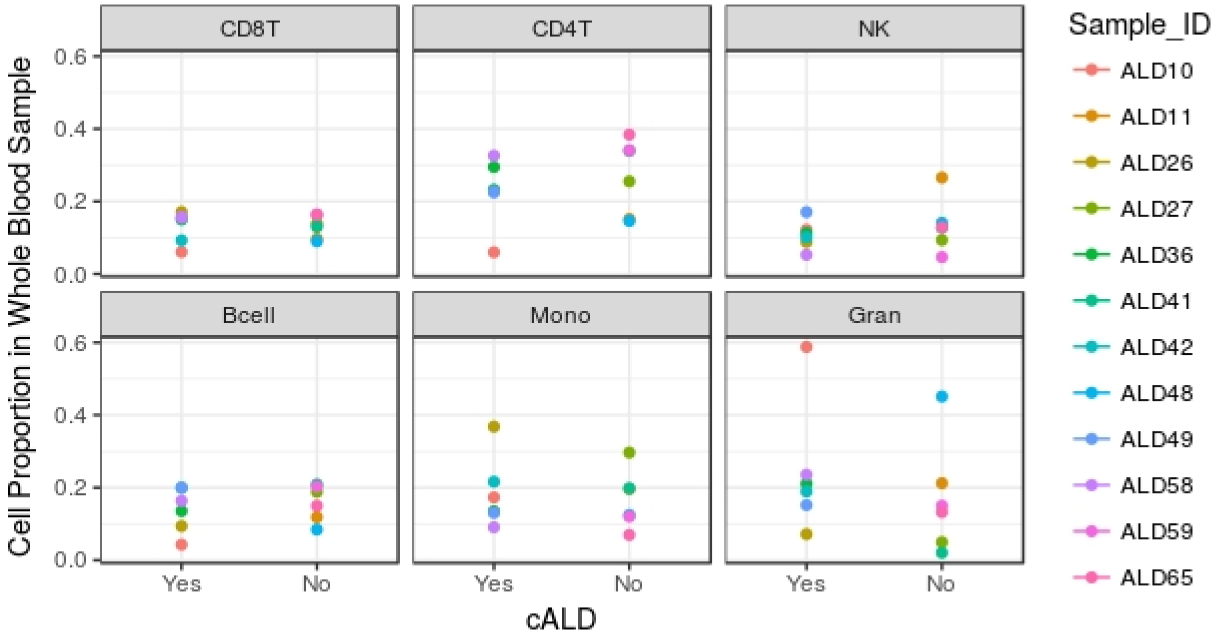

Figure 9 White Blood cell prportions estimated in each individual in this study.

Removal/Filtering of invariable probes in blood (NOT PERFORMED)

Following blood cell deconvolution procedure, we usually remove probes that are deemed invariable in blood samples (*Edgar et al., 2017*) to reduce penalty from multiple-test corrections and increase computing power.

Under request of the study, this step is not performed.

2. Differential Methylation Analysis

Linear Modeling with LIMMA

Due to the limited power in this study, we opted for LIMMA, a linear modeling package, to identify CpG loci that display significant changes in DNA methylation levels (beta values). This is our first approach - we will also identify DMRs (see below).

Limma was run with the following model:

design <- model.matrix(~**cALD+Age+Sib__Pair**, data = meta__ALD)

where Sib__Pair and cALD are kept as factors.

Figure 10 shows the Nominal P-value distributions following LIMMA. This is not especially surprising given the lower number of samples and the reduced power.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

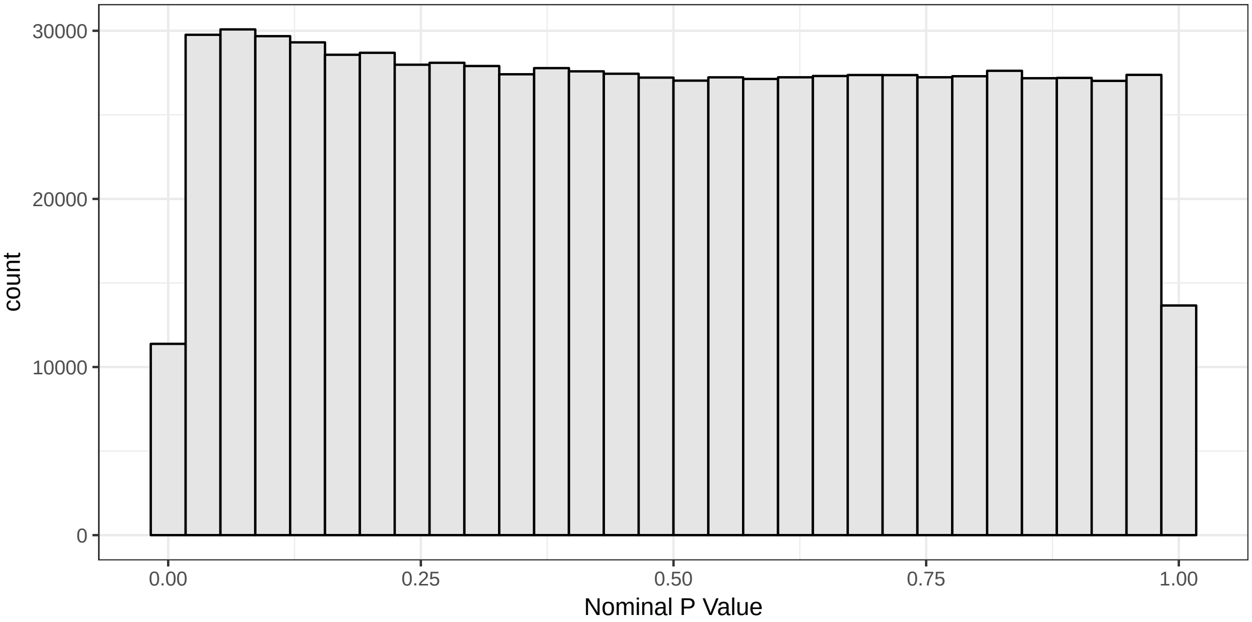

Figure 10 P-Value distributions of the LIMMA results for the x-ALD pilot.

Using a nominal P-value cut off of 0.0005, with a delta beta cut-off of 0.05, a Volcano plot was generated (**Figure 11**) to illustrate our Limma results.

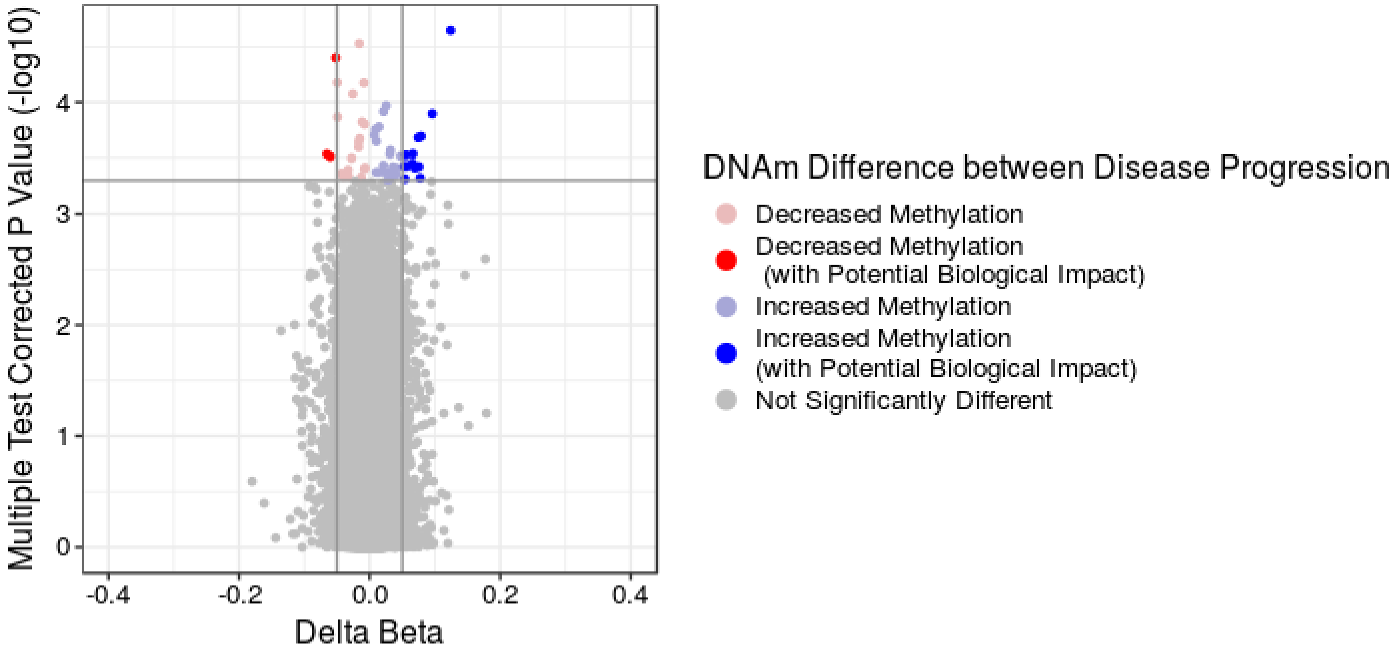

Figure 11 Volcano plots showing DNA methylation trends/Nominal P. Value of all CpGs examined by LIMMA.

The Volcano plot demonstrates that there are 13 hypermethylated and 3 hypomethylated loci that meet our cut-off. Note prior to Delta Beta filtering, 57 CpG Loci passed the nominal P of 0.0005. The list of these CpGs is shown below in Table 3.

Table 3 LIMMA Hits summary

	CpG	Chr	Coordinate	Nominal P. Value	Delta Beta	UCSC Gene Name
5	cg03329597	3	108125523	0.0000225	0.1241363	MYH15
43	cg20327444	13	110959647	0.0000296	-0.0158066	COL4A1;COL4A2;COL4A2
12	cg06134980	15	50551582	0.0000398	-0.0516862	HDC;HDC
27	cg15453257	15	86021188	0.0000664	-0.0494263	AKAP13;AKAP13
28	cg15692972	2	111751981	0.0000668	-0.0086545	ACOXL
23	cg12740438	5	1887364	0.0000842	-0.0256609	CTD-2194D22.4;IRX4;IRX4;IRX4;IRX4
47	cg21280320	1	152162025	0.0001077	0.0254437	
34	cg17754500	3	99224966	0.0001213	0.0214376	
9	cg03938532	15	55133091	0.0001271	0.0960255	
53	cg23422866	8	143326234	0.0001360	-0.0489825	TSNARE1
20	cg11183632	20	21503152	0.0001504	-0.0113722	
17	cg10076560	Y	21729144	0.0001576	-0.0077811	CYorf15A
38	cg18811130	10	131268886	0.0001666	0.0145092	MGMT
4	cg02911909	5	14992937	0.0001756	0.0090476	
13	cg06582708	17	21029295	0.0001975	0.0072062	DHRS7B
2	cg01927686	12	96617777	0.0002023	0.0789835	ELK3
44	cg21003497	10	81065938	0.0002079	0.0745339	ZMIZ1
22	cg11889808	7	157483448	0.0002117	-0.0151346	PTPRN2;PTPRN2;PTPRN2
1	cg01323954	15	83654524	0.0002243	0.0103706	FAM103A1
14	cg07965822	9	101610575	0.0002283	-0.0166617	GALNT12
30	cg15841063	1	63789500	0.0002533	-0.0168341	FOXD3
48	cg21899558	7	661634	0.0002695	0.0319282	PRKAR1B;PRKAR1B;PRKAR1B;PRKAR1B;PR
35	cg18397726	14	93395464	0.0002906	0.0663503	CHGA
11	cg04933168	22	36960937	0.0002917	0.0312946	CACNG2

	CpG	Chr	Coordinate	Nominal P. Value	Delta Beta	UCSC Gene Name
10	cg04674762	14	52487120	0.0002918	-0.0653594	NID2
32	cg16571642	7	158045996	0.0002973	0.0556382	PTPRN2;PTPRN2;PTPRN2
19	cg10473311	7	158046358	0.0003041	0.0477842	PTPRN2;PTPRN2;PTPRN2
56	cg27519373	19	57350292	0.0003068	-0.0601893	PEG3;PEG3;ZIM2;PEG3;PEG3;ZIM2;ZIM2
15	cg08376643	10	64880766	0.0003175	-0.0275281	
40	cg19181528	20	59542589	0.0003531	0.0657289	
39	cg19123882	11	63438084	0.0003668	0.0211877	ATL3
55	cg27200869	7	158045980	0.0003737	0.0532566	PTPRN2;PTPRN2;PTPRN2
7	cg03577632	16	85216169	0.0003738	0.0573334	
45	cg21158163	20	59542578	0.0003766	0.0756247	
42	cg20141969	15	50400942	0.0003782	0.0356483	ATP8B4;ATP8B4;ATP8B4;ATP8B4
46	cg21158431	6	167786059	0.0003838	-0.0069804	
36	cg18650367	10	133909949	0.0003862	0.0414542	
37	cg18717044	3	127401333	0.0003870	0.0694024	
54	cg25785281	1	197169849	0.0003973	-0.0079952	ZBTB41
26	cg14471191	10	117818509	0.0004022	-0.0328830	GFRA1;GFRA1;GFRA1
49	cg22240515	13	28558413	0.0004213	0.0229634	PRHOXNB
51	cg22618509	20	10385879	0.0004230	0.0101844	MKKS;MKKS
6	cg03400443	2	74347616	0.0004237	0.0142129	
57	cg27525902	15	49716247	0.0004264	0.0322684	C15orf33;FGF7
41	cg19458741	5	31268955	0.0004327	-0.0428216	CDH6
29	cg15819924	7	157886456	0.0004397	0.0397616	PTPRN2;PTPRN2;PTPRN2;PTPRN2;PTPRN2
52	cg23284931	11	13983273	0.0004400	-0.0422415	SPON1
16	cg09926783	15	49716645	0.0004449	0.0400268	FGF7;C15orf33
31	cg16514212	7	157530897	0.0004555	-0.0368928	PTPRN2;PTPRN2;PTPRN2;PTPRN2;PTPRN2
33	cg17426568	6	170418587	0.0004592	-0.0322089	
8	cg03651525	8	27630438	0.0004609	-0.0399234	CCDC25;CCDC25;CCDC25;CCDC25;CCDC25;CC
25	cg13900737	9	33755121	0.0004691	-0.0128809	PRSS3
21	cg11189868	2	239342095	0.0004769	0.0774915	ASB1
50	cg22403534	11	24284034	0.0004830	0.0303553	
3	cg02253142	15	52048211	0.0004860	0.0541157	TMOD2;TMOD2
18	cg10377756	12	92691863	0.0004958	-0.0116152	
24	cg13293729	6	28911902	0.0004981	0.0278667	

The 16 CpGs that pass the 5% Delta Beta are graphically represented in **Figure 12**. Unfortunately, it would seem like Family 6 is driving most of the effects?…Further analyses will follow

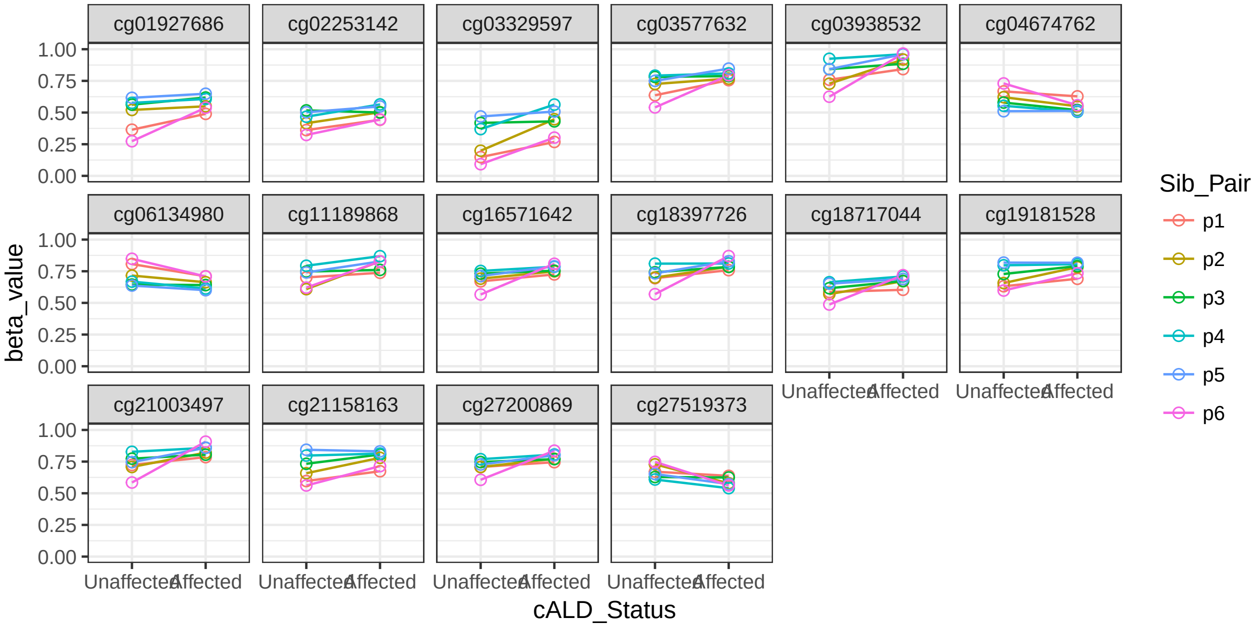

Figure 12 Delta Beta (Methylation) differences of the top Limma hits between cALD groups

Differential Methylation Region (DMR) Analyses: DMRCate

To assess DMRs between individuals at different disease progressions, we used the package DMRCate to pull out regions using cutoffs: 0.1 FDR and 5% Methylation change. Using this threshold we were able to pull out 99 CpGs. Among the top regions are PTPRN2 and HCG4P6, which are shown here. Interestingly though it would appear that…someone is driving the effect again.

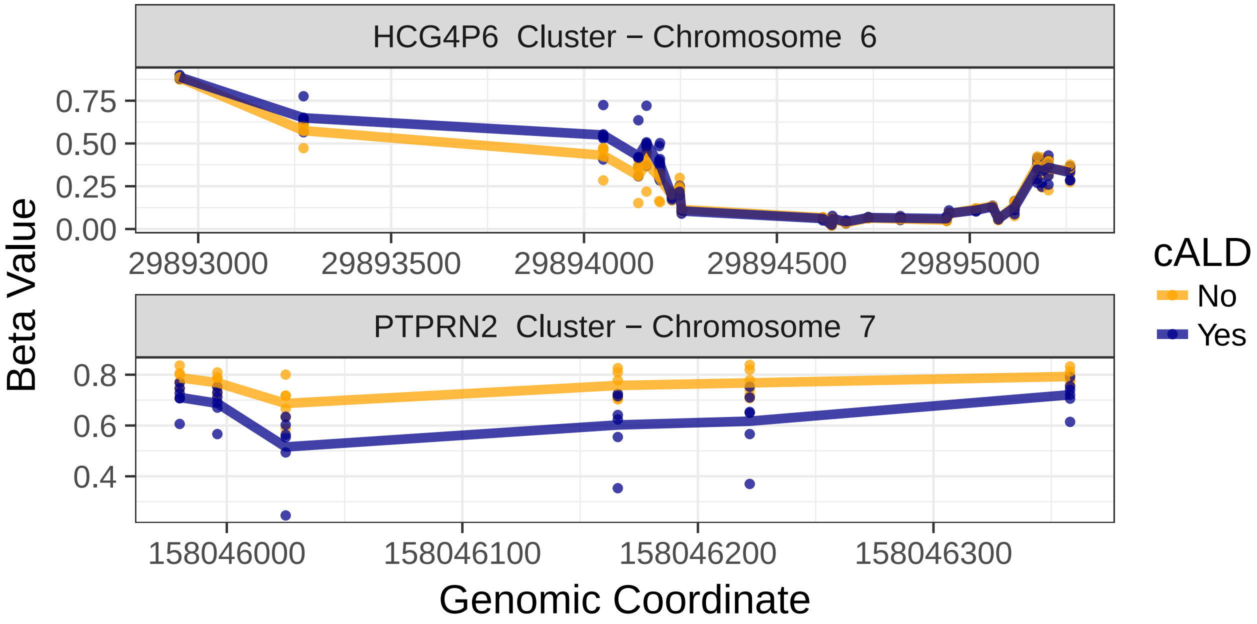

Figure 13 Representations of the top DMRs in the x-ALD pilot study.