Progress Report: Multiomic analysis of DNA, RNA and epigenomic networks for prognostication and novel target identification in Waldenstrom's Macroglobulinemia

Zachary R. Hunter, PhD & Steven P. Treon, MD, PhD

Lay Summary

We are pleased to update the committee on our progress to date. Since the last update we have greatly increased the stringency of our quality assurance filters and bioinformatic corrections to account for differences in reagents and sample handling. Ultimately this resulted in the removal of 25 of the original 302 samples for the study. In addition, we moved 11 samples from the tumor-normal paired sequencing group to tumor only sequencing due to extremely high levels of circulating tumor cells that resulted in the contamination of the control samples for these patients. All these efforts have resulted in a greatly improved data for all aspects of the study, including the whole exome sequencing, and RNA sequencing studies. This cleaned data has allowed us to perform mutational driver analysis, identify novel mutated genes, perform mutational enrichment studies and integrate/cross check results between data sets, all to make new observations that were not previously possible. Having this finalized has allowed us to move forward with the analysis of the clinical data as well which includes a median follow up time of 7.3 years and the major clinical events that took place during this time. This the most detailed clinical data set assembled by the Bing Center to date. This improved data has also been incorporated into our multi-data set network analysis that aims to study transcription factor – gene relationships that drive WM. To date we have constructed three networks consisting of one that includes all MYD88 mutated WM, one that is only CXCR4-mutated, and one that is only CXCR4 unmutated. These are being analyzed and contrasted with each other and a previously published network based on healthy B-cells that have been immortalized for laboratory analysis using the Epstein-Bar virus (EBV). Early results look promising and are corroborated by observed changes in the epigenetic data sets that measure the ability of transcription factors to bind to DNA. We have made detailed plans with our collaborators on how to proceed with the network and artificial intelligence driven genomic subtyping analysis for the reminder of the study and plan to have this data available by the end of the study period. Our final specific aim working with bone marrow pathology slides is running behind due to heavy employee turn-over throughout the Harvard cores and with our collaborators. We do have the first cohort of slides sectioned and ready for staining that will allow us to train the AI system, but more will be needed for the full WM subtype analysis. As we are not the only project depending on these slides, we are pushing hard to get this aspect of the project back up to speed. In conclusion, all the work we have put into coordinating and curating this large project has started to pay off. We hope these efforts will lead to important insights into the biology of WM and create a publicly available biological road map for understanding WM signaling that can be used to drive future research and therapeutic development.

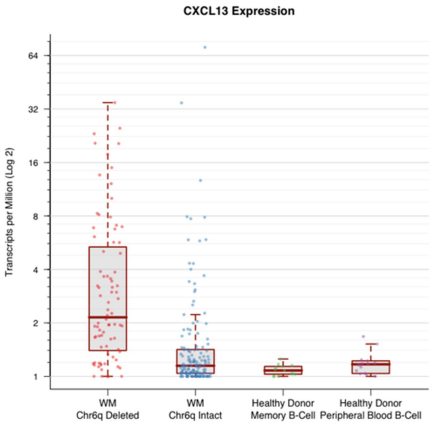
Introduction

This proposal's objective is to integrate data from existing DNA, RNA, and epigenomic sequencing studies derived from the same biopsies to define molecular subtypes and discover novel signaling networks in untreated patients with Waldenstrom's Macroglobulinemia (WM). Building on our previous roadmap grant, *Transcriptional Characterization of Untreated Patients with Waldenström's Macroglobulinemia,* we have assembled a next generation sequencing dataset of 305 whole exomes (WES), 274 RNA Sequencing studies (RNASeq), 100 methylation studies using enhanced reduced representation bisulfite sequencing (ERRBS), 41 transposase-accessible chromatic studies (ATAC-seq) and 40 5-hydroxymethylcytosine studies. All of the sequencing is now complete, and characterization of the individual data sets is well underway.

Our central hypothesis is that many signaling networks cannot be inferred through isolated study of mutation, transcription, and epigenetic differences, but when taken together, signaling cascades critical to the growth and survival of WM can be observed. This is particularly pertinent as mutant *MYD88* mouse models have demonstrated that *MYD88* alone is not sufficient for WM pathogenesis. Using these parallel data sets we aim to improve our understand of signaling networks and genomic subtypes of WM and use this information to develop survival and therapeutic response prognostic tests as well as discover novel drug targets. A part of this project includes the use of exploratory immunohistochemistry staining and machine learning analysis to develop proxy marker test that could be run in standard pathology laboratories without specialized equipment. As is often the case with these large projects, there is a lot of upfront work that goes into data collection, quality assurance and analysis before things start to pay off. I am pleased to say that things are at the stage where it is all coming together very quickly, and we provide some of the highlights below.

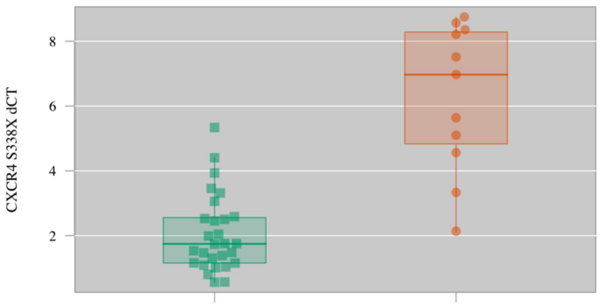
Specific Aim 1: To characterize the genetic subtypes of untreated Waldenstrom's Macroglobulinemia

Sub Aim 1-1: Curation of the multiomic data set and clarification of ambiguous events: Important breakthroughs have made in nearly every aspect of the data. The largest of these being the implementation of a more aggressive approach to batch correction in the RNASeq data using ComBat-Seq and using variance stabilizing transformations from the DESeq2 package on the batch corrected count data for downstream correlation and clustering analysis. This has greatly improved the quality of the RNASeq data and has given us greatly improved results, including our PANDA network analysis. One small example is the strong association of chr6q deletions with increased *CXCL13* expression that can be observed in the new data. Based on our previously published work on *CXCL13*, we would expect Del6q to be associated with increased LPC infiltration of the bone marrow. Indeed, in this data set the median involvement was 60% (range 4-95%) when chr6q was deleted compared with a median of 40% (range 5-95%; p=0.0006) for chr6q intact patients. The transformed data for correlations and differentially expressed gene list have been integrated into our custom ATACSeq peak annotation package which greatly aided our interpretation of the differential peak data and allowed for better integration with our PANDA network analysis findings which will be discussed in the network analysis section. The WES data set has matured significantly and has been thoroughly pruned of samples failing QC or having ambiguous/competing diagnosis that may complicated the analysis. The final sample breakdown of the project is shown in the table below though each data set does have several unique/non-overlapping WM samples that are not shown here.



Final WES Cohort Overlaps with Other Sequencing Data

Group	N	RNASeq	ERRBS	ATACSeq	Original_WGS
Follow_Up_Pairs	6	6	2	2	0
Paired	244	218	37	34	32
Unpaired	30	26	4	1	4

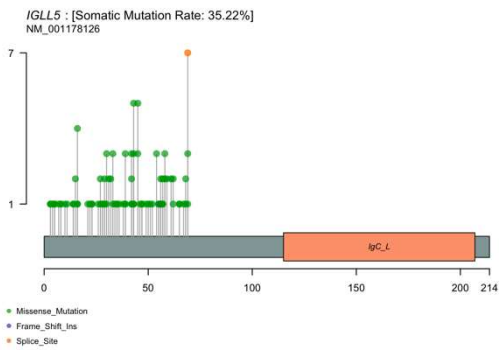


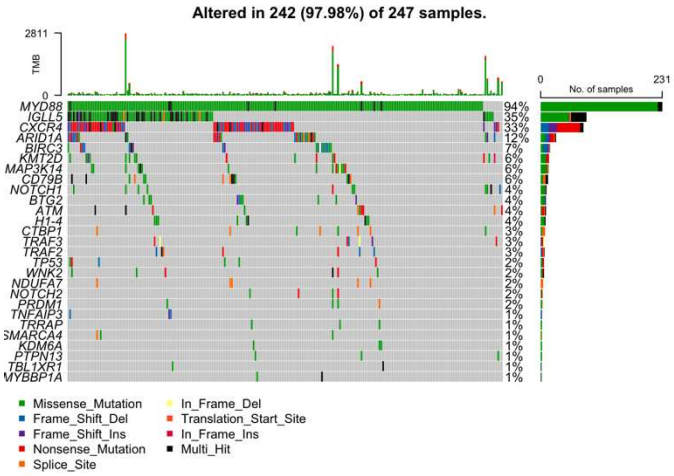
CXCR4 Mutations missed by WES were often very subclonal with allele fractions ≤ 3%

gene_name	n_syn	n_mis	n_non	n_spl	n_ind	qglobal_cv
MYD88	0	237	0	0	6	0.0000000
CD79B	1	13	0	9	0	0.0000000
CXCR4	0	1	48	0	28	0.0000000
ARID1A	1	13	8	2	10	0.0000000
BIRC3	0	6	1	0	11	0.0000000
H1-4	3	13	1	0	0	0.0000001
BTG2	2	7	0	0	3	0.0000052
TRAF3	0	3	3	0	4	0.0000301
NDUFA7	0	0	0	5	0	0.0001078
TRAF2	0	2	2	1	3	0.0001148
TNFAIP3	1	0	0	0	5	0.0042909
CTBP1	1	3	0	5	0	0.0214182

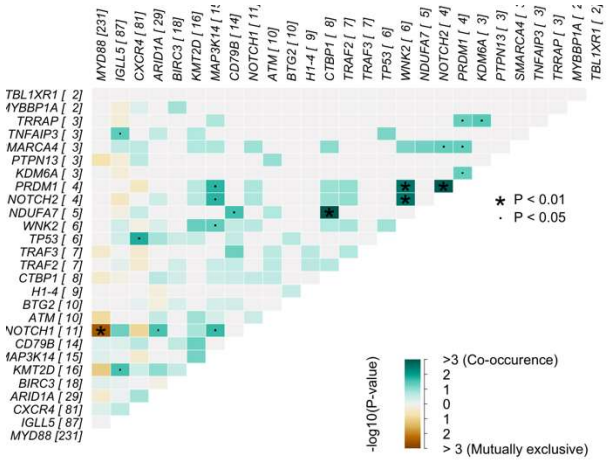
We can use our lab-based results for CXCR4 and MYD88 for comparison see how the sequencing performed. A total of 225 of the 238 (94.5%) positive *MYD88* cases from the lab based AS-PCR assay were found by WES. All 4 of the *MYD88* wild-type calls samples were correctly identified. The AS-PCR dCT values of the concordant samples was 2.14 (range -0.6 - 6.97) while the samples missed by the WES analysis had a median dCT of 2.91 (range 1.84 - 8.09; p = 0.00042). Overall, this gives us a sensitivity of 94.8% and a specificity of 100%. The false negative results with low dCT were often caused by localized capture failures that were not representative of the sample as a whole and resulted in single digit coverage of the p.L265P region. As our ability to test for mutations in *CXCR4* is depends on Sanger sequencing for all but the p.S338 nonsense mutations, we fully expected to pick up some novel cases in the WES data while missing some subclonal S338 mutations, which was in fact the case. Overall, 107 WHIM syndrome like mutations were observed in 247 (43.3%) of patients. The lab based and WES analyses found mutations in 102 (95.3%) and 80 (74.8%) patients, respectively. This results in a sensitivity of 95.5% for lab based and 79.9% for WES methodologies. For CXCR4, the discrepancy was driven largely by subclonal variants as shown in the graph above. While imperfect, this is by far the most sensitive large scale NGS result to seen to date. We are now able to move forward with a high degree of confidence that these samples have been fully screened clinical and informatically for data quality.

Using mutational background measurments we can look for genes that exhibit mutation patterns that appear to be under selection preasure and thereby likely candidates for driver mutations. The results of this analysis using the R package *dndxcv* from the Sanger Institute with Hg38 covartiates is shown here. While this mostly reaffirmed the published literature, there are several novel gene candidates as well. It is worth mentioning that *IGLL5* was featured prominantly prior to the the introduction of the covariate matrix with missense mutations found in 35% of the samples. With the covariate data *IGLL5* was deemed not signifcant and it is notable that the mutation pattern is overwhelmingly silent (not shown) and all of it is in the beginning of the gene consistent of what is known about somatic hypermutation off targeting. Even so, we are continuing to analyse the role of *IGLL5* mutations in WM. Even if they prove to be a biological by-product, their presense could be a indirect marker for relevent information like time spent in the germlinal center.



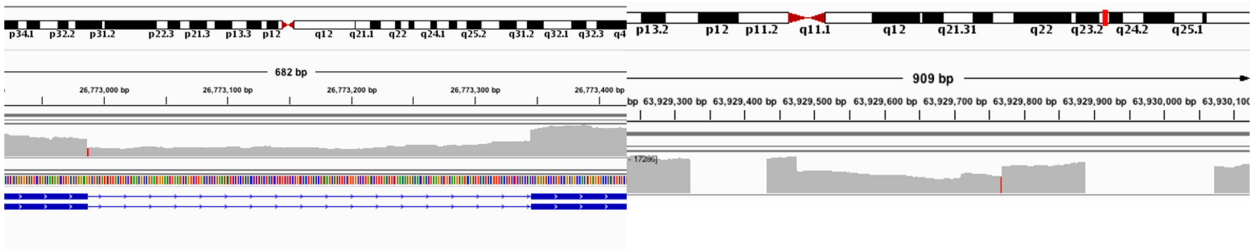


Sub Aim 1-2: Molecular subtype identification: While we are still optimizing some of the structural variant analysis and cancer cell fraction calculations which may impact the clustering, we have started to look at mutational patterns in the WES small variant data as shown below. We have begun work on both the standard methodologies such as clustering and NMF but have also prepped the mutational and RNASeq data for more advanced message passing co-clustering algorithms such as similarity fusion networks (a variation of which is also being used in the PANDA analysis). Even at this early stage there are some interesting patterns such as *NOTCH1*/*MYD88* mutual exclusivity and the *TP53/CXCR4* co-occurrence. As soon as the full data set is integrated, these efforts will begin in earnest.

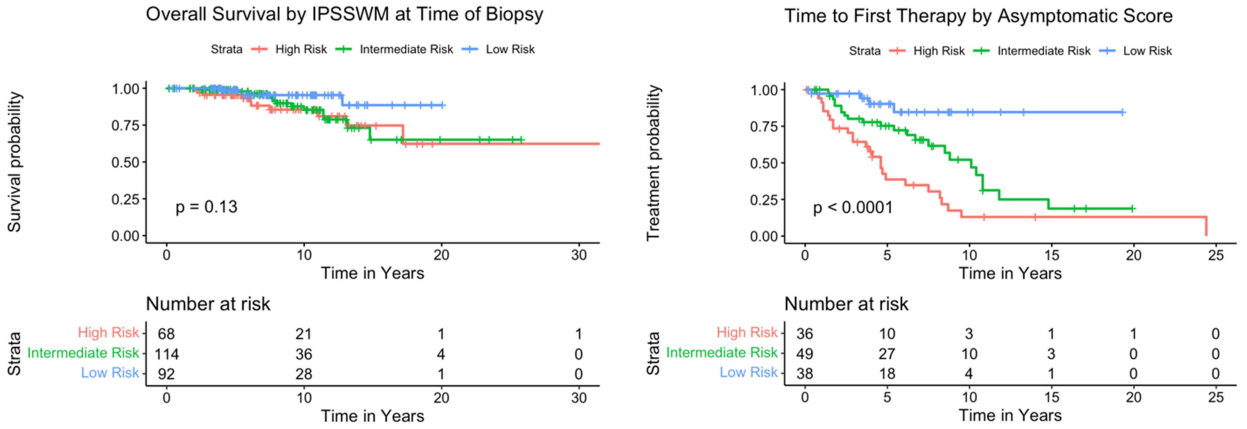


Dndxcv driver gene analysis of the WES data with genes ranked in order of driver probability.

Sub Aim 1-3: Multiomic characterization of genomic subtypes: As mentioned earlier, we have fully integrated the ATAC data with WES and RNASeq. Likewise, we are using the RNASeq data to interpret ambiguous WES findings such as IGLL5 as well as validate putative splicing mutations such as the ones shown below. We are also using the ERRBS and ATAC data to help us interpret the PANDA networks. The full multiomic subtypes will depend on the completion of the NMF and the co-clustering neural network analysis such as the similarity fusion and/or affinity propagation results that should be available this spring.



Sub Aim 1-4: Multiomic characterization of clinical subtypes: Similar to the previous sub aim, we are waiting on the structural variant data integration, but we have already gone ahead to characterize the clinical data set. With a median follow up of 7.3 years (range 0.2 - 31.9 years), only 28 of the 260 (9.72%) patients have died for whom follow up data was available. This is really a testament to the massive improvements in patient care and therapeutic options during this period. The IPSS stratification did not perform particularly well for this data set. This may be due to the introduction of novel therapeutics after the system was developed. With 210 patients having gone on to receive some type of therapeutic intervention we can call the median time to first therapy at 2.75 years for this data set. The asymptomatic scoring system for developed by Bustoros et al. in 2019 did an excellent job of predicting time to first therapy in asymptomatic patients as shown below.



We have also begun to perform clinical characterizations of the mutations found in the study. Here below we show the most significant data for CXCR4 mutation status as an example.

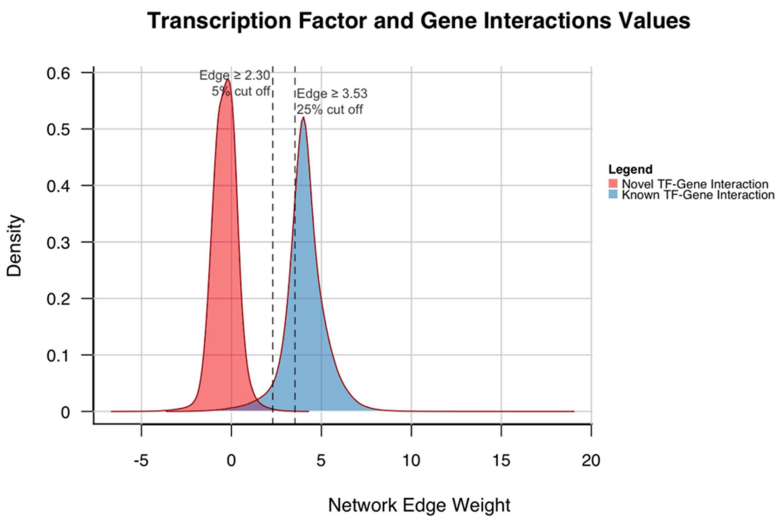
Name	CXCR4 Wild Type	CXCR4 Nonsense	CXCR4 Frameshift	WT vs. NS p.value	WT vs. FS p.value	NS vs. FS p.value
alymph	1.730	1.625	1.535	0.2400958	0.0504510	0.4620853
aneut	3.935	3.615	3.720	0.0647219	0.2923380	0.5125728
iga	56.000	61.000	39.000	0.1934252	0.0959665	0.0040236
igg	643.500	572.000	491.000	0.3201640	0.0601286	0.3544043
igm	2,554.000	4,087.500	4,104.000	0.0130051	0.0024894	0.8684056
lymphocytes	24.000	30.000	28.000	0.0966247	0.2654718	0.6940872
plasmacells	1.000	2.000	3.000	0.0460366	0.0059510	0.4225291
plt	261.000	197.000	236.500	0.0000876	0.0238686	0.1816263
timewmtobmbx	0.800	0.900	2.800	0.7510049	0.0544448	0.1035890
wbc	6.675	5.900	6.190	0.0237993	0.0222521	0.9787547

Name	CXCR4 Wild Type	CXCR4 Nonsense	CXCR4 Frameshift	WT vs. NS p.value	WT vs. FS p.value	NS vs. FS p.value
Adenopathy	67/165 (40.6%)	8/54 (14.8%)	6/43 (14%)	0.0004588	0.0010693	1.0000000
CD10	17/171 (9.9%)	2/59 (3.4%)	1/50 (2%)	0.1692978	0.0817773	1.0000000
CD23	46/171 (26.9%)	23/59 (39%)	17/50 (34%)	0.0993401	0.3739680	0.6909052
Mast Cells Present	129/171 (75.4%)	51/59 (86.4%)	37/50 (74%)	0.0990309	0.8535369	0.1432710
WM Specific Familial History	7/116 (6%)	6/48 (12.5%)	6/38 (15.8%)	0.2043126	0.0881133	0.7582932

Specific Aim 2: To conduct the first multiomic network analysis of Waldenstrom's Macroglobulinemia

Our collaboration with the T.H. Chan Harvard School of Public Health Quantitative Biomedical Research Center is very active now and we are taking an iterative development approach for optimizing some of these analyses. There is a lot more happening than can be covered here, so we are presenting the current scope of work with a few specific examples to ground the narrative.

Sub Aim 2-1: Network discovery and quantification with EQTL, PANDA and LIONESS: After evaluating our initial PANDA/LIONESS networks, we realized that that we needed to be more aggressive in filtering low expressed genes and switch to making hard corrections for batch effect up front, rather than simply incorporating the information into the overall model to compensate. We have now created three network models. The first is for all *MYD88* mutated WM while the other two are limited to *CXCR4*-mutated and *CXCR4*-Wild-Type WM, respectively. We have obtained a related PANDA network generated from 132 EBV transformed B-cells from the Gene Regulatory Network Database (GRAND) to use as a B-cell background comparison. These networks are 20,130 by 645 matrices documenting transcription factor – gene interaction z-scored likelihood. Known interactions are incorporated as a part of the deep learning model allowing us to stratify the results by known vs. novel interactions. It goes without saying that these are complicated data sets to analyze and interpret, but there are several established approaches. We can compare edge weights between two networks and use the inverse cumulative distribution function to look at the top 20% of differential findings and limit the results to cases where at least one group had a believable interaction to get differential changes between networks. We can also look at overall connectedness of genes and transcription factors. This can be done summing up all the network edges per node or be looking at the total number of edges connecting to nodes that pass a given filter such top edge weights for novel interactions. Finally, since the data is already in a z-scored distribution with each transcription factor connecting to all genes, we can use gene set enrichment algorithms such as limma's Camera or the Broad's GSEA to compare the functional enrichment of a given transcription factor's interactions within and between networks. As one of the main drivers in interactions changes is epigenomic regulation, we can corroborate the PANDA findings with ATACSeq and ERRBS. To give a specific example, we ranked genes in the *MYD88* network by those with the most incoming edges in the top 1% of edge weights. It was a nice surprise to find that the top gene was *ATRN*, a gene that is upregulated in WM and promotes immune cell aggregation and chemotaxis during inflammation. A peak strongly associated with *ATRN* was also one of our top ATACSeq findings in our WM versus healthy donor B-cell analysis. We then checked to find overlaps between the top transcription factors connected to *ATRN* and motifs in these differentially regulated chromatin regions



Distribution of network edge weights representing the likelihood of a particular transcription factor – gene interaction stratified based on known versus novel relationships from the *MYD88*-Mut WM network.

Genes sorted by number of edges in the top 1%

Gene	N_Edges	Symbol
ENSG00000088812	13	ATRN
ENSG00000224914	13	LINC00863
ENSG00000134748	12	PRPF38A
ENSG00000171311	12	EXOSC1
ENSG00000183155	12	RABIF
ENSG00000075975	11	MKRN2
ENSG00000119004	11	CYP20A1
ENSG00000130731	11	METTL26
ENSG00000152475	11	ZNF837
ENSG00000156384	11	SFR1

to demonstrate overlapping signals.

The last planned analysis for the current set of PANDA networks is to use similarity fusion networks to try to identify smaller, more self-contained sub-networks in the larger one. If we can identify relevant regulatory functions for these subnets, it will be an interesting finding in its own right, but also can be used for dimensional reduction of the LIONESS data where we can score each sample by the relative strength of these subnetworks and analyze the resultant scores. Once we feel the PANDA level analysis is mature, we will progress on to LIONESS to generate a custom network for every study sample allowing us to investigate smaller populations with a greater degree of nuance.

Sub Aim 2-2 Network functional evaluation: While still in the early stages, we have begun using the ATACSeq and ERRBS data to confirm and validate the predicted transcription factor interaction changes emerging from the PANDA network analysis. We are also processing the ATAC data be optimized for investigating transcription factor foot printing analysis (transcription factor binding interferes with the transposon activity making a small divot in the coverage peak corresponding to the binding site). Wet lab validation will take place once the top results have been identified.

Specific Aim 3: Assess multiomic Impact on the WM Microenvironment and develop proxy assays for subtype and network activation status

This specific aim continues to be plagued with delays due to the large amount of employee turn-over in our core facilities. However, progress is being made and we have several new hires at the Bing Center that will be starting next month who will be able to assist, particularly with the CyTOF analysis.

Sub Aim 3-1: CyTOF analysis of WM and microenvironmental networks: This sub aim is presently on hold. While the samples are prepared, the departure of Guang Yang for private industry halted progress. We have a new bioinformatic analyst and a new senior wet lab scientist starting next month which will allow this aim to get back on track.

Sub Aim 3-2: Quantitative biomedical imaging of pathology slides: After several setbacks due to key personal in pathology retiring or leaving for other positions, we have been able to section most of the slides. We lost 4 key collaborators since this project began but we have continued to push hard to get this done as there are several important studies depending on this data including Ruben Carrasco's Roadmap grant and Maria Luisa's Kyle award project. We hope to start QCing the cut slides next week before sending them for staining at the Brigham clinical pathology facility. Once stained, they will be imaged by Brigham research pathology and sent to myself and Brian Lawney to begin our AI driven image analysis. The use of three different cores for cutting, staining, and imaging, respectively, was unfortunately necessary due again to loss of key personnel in their facilities as well. One additional change from the previous update was a switch from CD19 to PAX5 as the third stain supplementing the H&E and Giemsa slides. This was due to the relative performance of the two antibodies in IHC and the advantage that PAX5 should not be expressed in lymphoplasmacytic cells and is well expressed in WM which may aid the AI in distinguishing normal vs WM cells based on PAX5 expression and morphology.

Sub Aim 3-3: Development of assays indicative of WM subtype membership: This aim is on hold pending the finalization of the multi-omic subtype and network analysis. This is

the final step in the project where we try to identify proxy markers to infer group membership and/or network activation status.