UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

HTT CHIPSEQ ANALYSIS
WT & Q111het & QQ

Amol Carl Shetty

September 2, 2021

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Dataset Descriptions

- Dataset 1 (GSE102750)

Samples	11	
Conditions	Wt ChIP	3 replicates
	Wt_Input	2 replicates
	Q111het_ChIP	4 replicates
	Q111het_Input	2 replicates
Matched ChIP-Input samples	No	
# Sequencing Runs / sample	2	
Sequencing Read Length	25	
Tissue	Striatal	

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Dataset Descriptions

- Dataset 3 (Genewiz :: 30-415606808)

Samples	12	
Conditions	Wt ChIP	3 replicates
	Wt_input	3 replicates
	QQ_ChIP	3 replicates
	QQ_input	3 replicates
Matched ChIP-Input samples	Yes	
# Sequencing Runs / sample	1	
Sequencing Read Length	150	
Tissue	Striatal	

REFERENCE-BASED ALIGNMENT

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Alignment Summary Statistics

- Reads assessed for quality and trimmed (if needed)
- Alignment of reads using Bowtie2 against mm10
- Alignment Summary for GSE102750

Htt Genotype	Experiment	# Total Reads	# Mapped Reads	%Mapped Reads	% Properly Paired
Q111het Rep1	ChIP Run 1	21,034,752	20,752,498	98.6582	97.9840
Q111het Rep1	ChIP Run 2	18,851,976	18,602,823	98.6784	97.9965
Q111het Rep1	Input Run 1	18,405,330	18,199,073	98.8794	67.9330
Q111het Rep1	Input Run 2	16,310,914	16,129,846	98.8899	68.2041
Q111het Rep2	ChIP Run 1	20,673,202	20,432,692	98.8366	97.6156
Q111het Rep2	ChIP Run 2	18,565,752	18,350,358	98.8398	97.6368
Q111het Rep3	ChIP Run 1	25,903,280	25,590,856	98.7939	96.8596
Q111het Rep3	ChIP Run 2	23,299,664	23,017,227	98.7878	96.8657
Q111het Rep3	Input Run 1	18,922,162	18,729,955	98.9842	80.2070
Q111het Rep3	Input Run 2	16,805,344	16,637,907	99.0037	80.3989
Q111het Rep4	ChIP Run 1	26,277,440	25,972,482	98.8395	97.0909
Q111het Rep4	ChIP Run 2	23,698,832	23,424,035	98.8405	97.1171

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Alignment Summary Statistics

- Reads assessed for quality and trimmed (if needed)
- Alignment of reads using Bowtie2 against mm10
- Alignment Summary for GSE102750 (… contd …)

Htt Genotype	Experiment	# Total Reads	# Mapped Reads	%Mapped Reads	% Properly Paired
WT Rep1	ChIP Run 1	17,190,838	17,015,897	98.9824	98.2361
WT Rep1	ChIP Run 2	15,354,050	15,195,906	98.9700	98.2541
WT Rep2	ChIP Run 1	23,313,776	22,593,809	96.9118	92.1647
WT Rep2	ChIP Run 2	20,958,948	20,376,596	97.2215	92.2698
WT Rep2	Input Run 1	14,547,116	14,373,126	98.8040	83.4925
WT Rep2	Input Run 2	12,942,328	12,787,878	98.8066	83.7313
WT Rep3	ChIP Run 1	27,288,086	27,002,947	98.9551	96.3493
WT Rep3	ChIP Run 2	24,647,984	24,391,732	98.9604	96.4145
WT Rep4	Input Run 1	15,852,484	15,661,043	98.7924	63.0058
WT Rep4	Input Run 2	14,081,646	13,914,829	98.8154	63.5632

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Alignment Summary Statistics

- Reads assessed for quality and trimmed (if needed)
- Alignment of reads using Bowtie2 against mm10
- Alignment Summary for Genewiz :: 30-415606808

Htt Genotype	Experiment	# Total Reads	# Mapped Reads	%Mapped Reads	% Properly Paired
QQ1	ChIP	65,775,652	65,251,350	99.2029	87.4658
QQ1	Input	53,344,668	52,780,560	98.9425	80.9578
QQ2	ChIP	82,480,558	81,821,505	99.2010	88.0959
QQ2	Input	66,439,270	65,891,851	99.1761	90.7416
QQ3	ChIP	78,929,922	78,323,098	99.2312	88.4246
QQ3	Input	67,013,246	66,483,133	99.2089	92.9582
WT1	ChIP	68,600,884	68,145,010	99.3355	86.828
WT1	Input	41,203,654	40,699,440	98.7763	87.5731
WT2	ChIP	34,938,936	34,608,664	99.0547	79.8049
WT2	Input	32,782,802	32,158,328	98.0951	84.1743
WT3	ChIP	62,514,584	62,007,057	99.1881	87.4798
WT3	Input	83,080,630	82,338,814	99.1071	90.3013

PEAK DETECTION

Moving ahead with …

~30M w/ Pooled Input

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Peak Calling (~30M w/ pooled input)

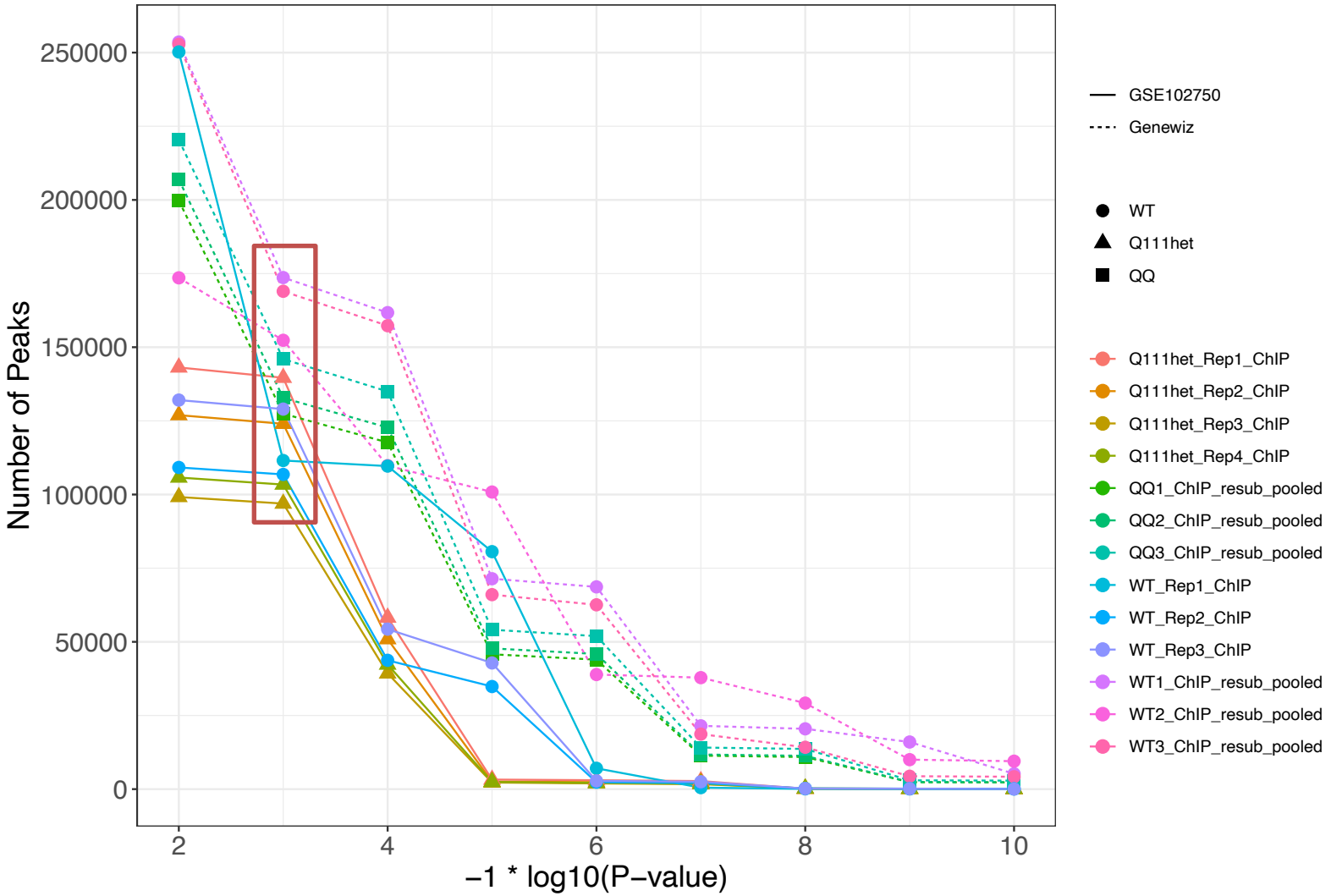
- Pooled inputs / genotype for GSE102750 and Genewiz
- Peak calling using MACS (--p-val 0.01, --scale-to-large)

Dataset	Genotype	# Peaks	# Filtered Peaks (p < 1e-3)
GSE102750	Q111het_Rep1	143,131	139,629
GSE102750	Q111het_Rep2	126,943	124,006
GSE102750	Q111het_Rep3	99,164	96,922
GSE102750	Q111het_Rep4	105,765	103,381
Genewiz	QQ1	199,752	127,429
Genewiz	QQ2	206,929	132,794
Genewiz	QQ3	220,383	145,986
Genewiz	WT1	253,585	173,641
Genewiz	WT2	173,556	152,350
Genewiz	WT3	252,807	168,985
GSE102750	WT_Rep1	250,230	111,541
GSE102750	WT_Rep2	109,185	106,825
GSE102750	WT_Rep3	132,059	128,998

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Peak Calling (~30M w/ pooled input)

- Pooled inputs / genotype for GSE102750 and Genewiz



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Reproducibility metrics
  - '+' implies ≥3 samples among samples belonging to selected group(s)
  - '-' implies <3 samples among samples belonging to selected group(s)
  - '*' implies group samples from selected group(s) were ignored

	2017.WT	2017.Q111	2020.WT	2020.QQ	# Peaks
LowReproducibility	<3	<3	<3	<3	38,900
Potential WT-specific	+	+	+	-	2,977
WT-specific	+	*	+	-	9,627
WT-mHTT-shared	+	*	+	++	1,656
Potential mHTT-specific	-	+	-	+	1,440
mHTT-specific	-	*	-	+	3,173
Q111-specific	-	+	-	--	253

Moving ahead with …

DIFFERENTIAL BINDING

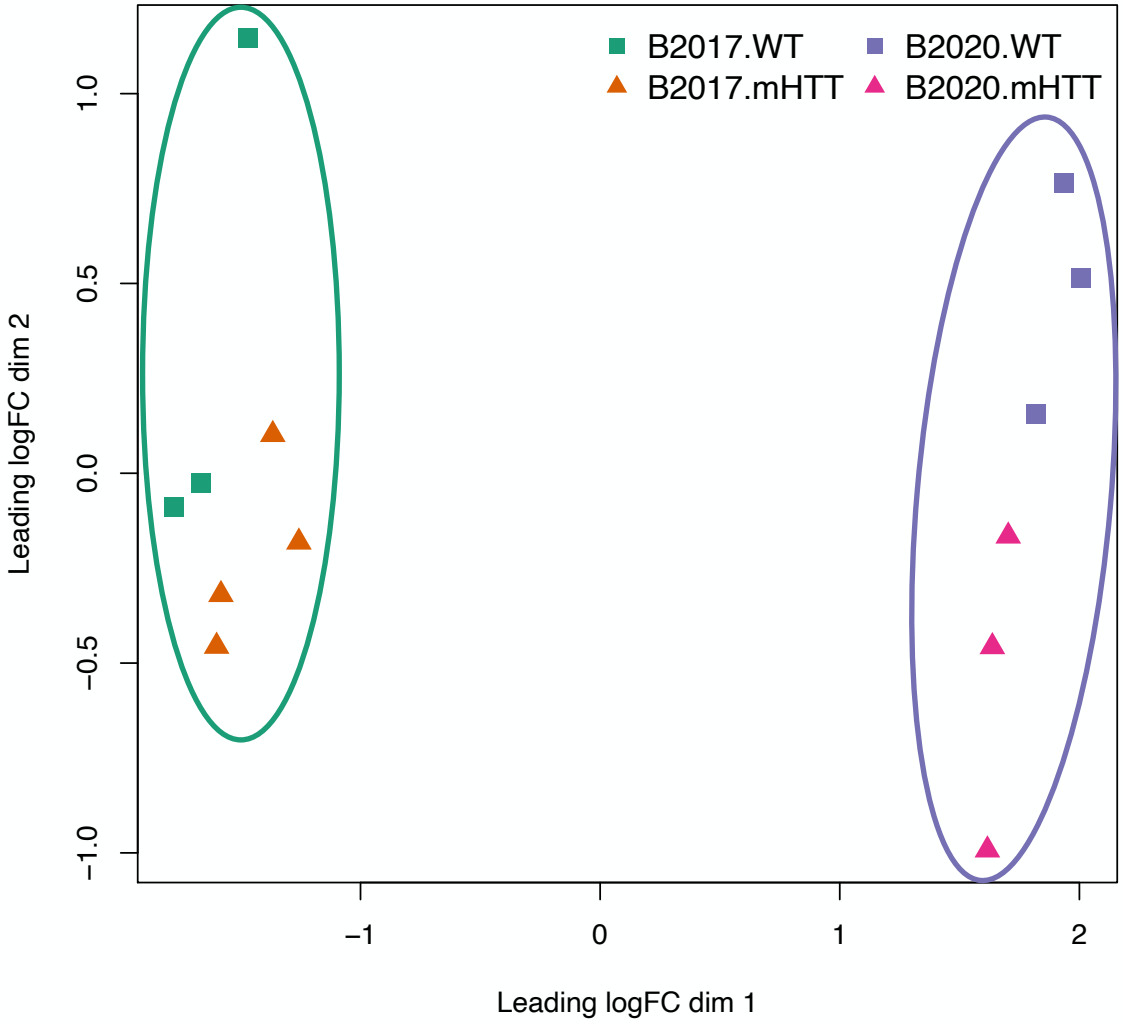
Version 3: p1e3 shared by ≥3 samples
(latest reproducibility groups)

BATCHES ANALYZED TOGETHER

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Sample clustering (B2017 and B2020 analyzed together)



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Differential binding between genotypes

Comparison	Significance Cut-off	# Regions LFC > 0	# Regions LFC < 0
2017 mHTT vs 2017 WT	P-value < 0.05	2,776	1,369
2017 mHTT vs 2017 WT	FDR < 0.05	10	4
2020 mHTT vs 2020 WT	P-value < 0.05	266	1,619
2020 mHTT vs 2020 WT	FDR < 0.05	0	0

?? More Q111het enriched peaks than WT enriched peaks in B2017 ??

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Differential binding between genotypes

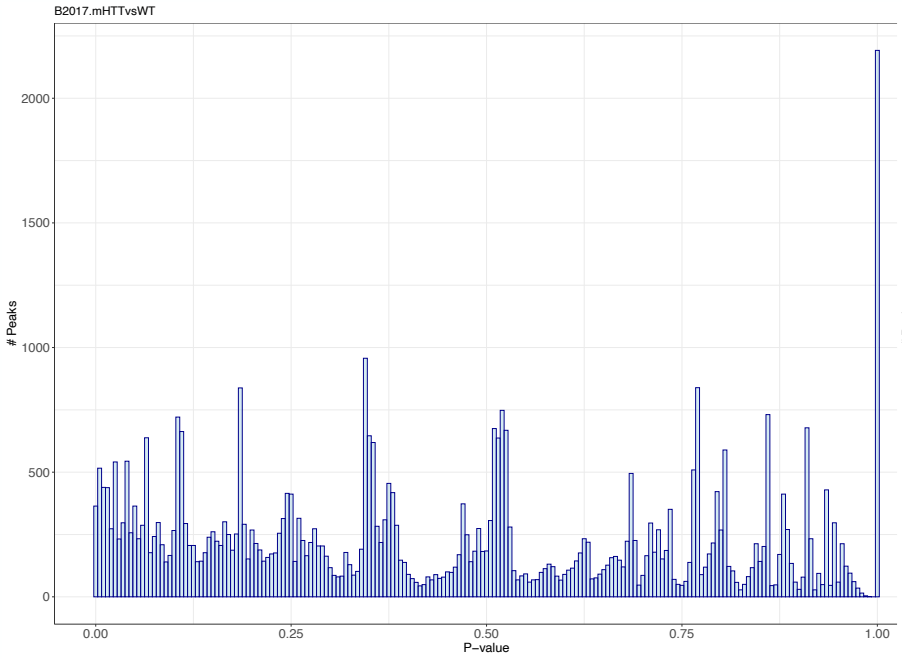
Htt_ChIPSeq.downsample_30000000.samples.diffbind.v20210831.HTT.diffbind_results.pdf

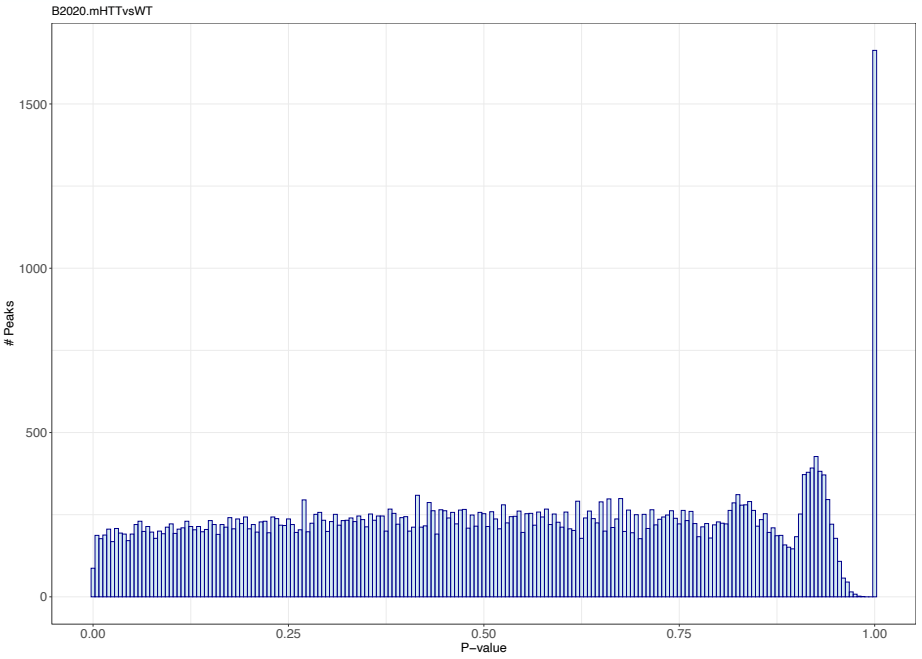
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value for each Peak Region

B2017 mHTT vs WT	B2020 mHTT vs WT



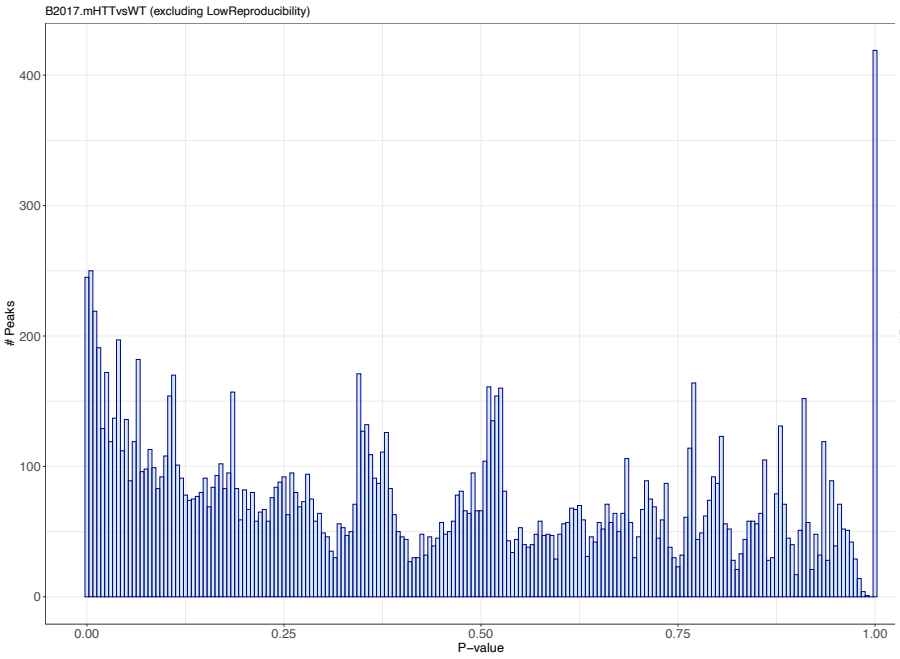


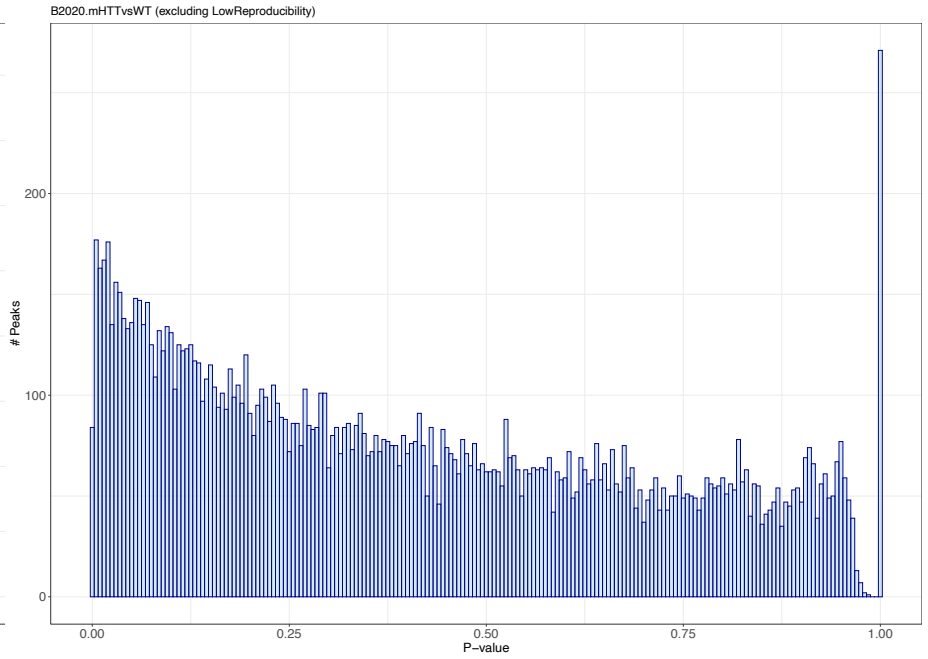
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value for each Peak Region (excluding LowReproducibility peaks)

B2017 mHTT vs WT	B2020 mHTT vs WT

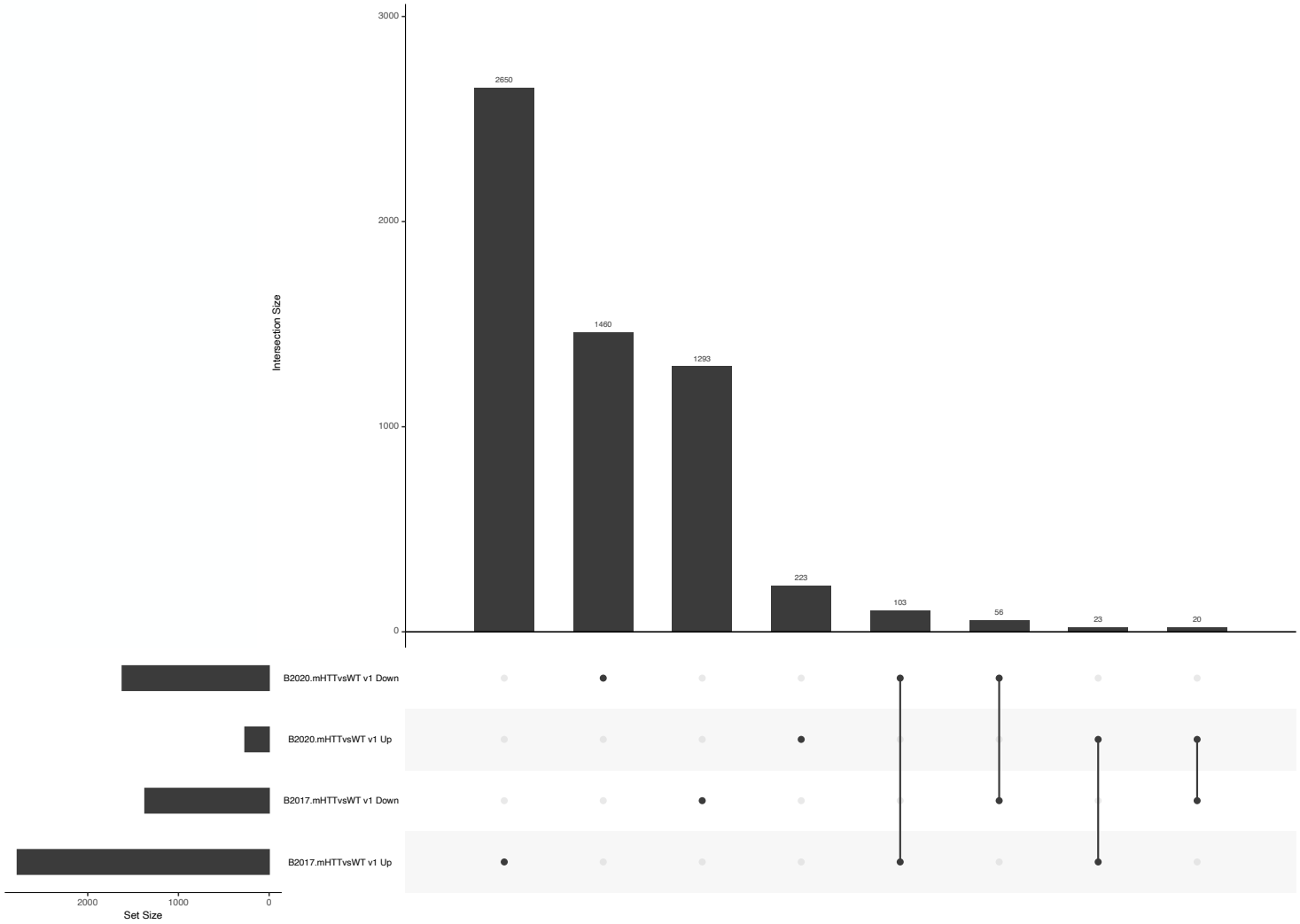




UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

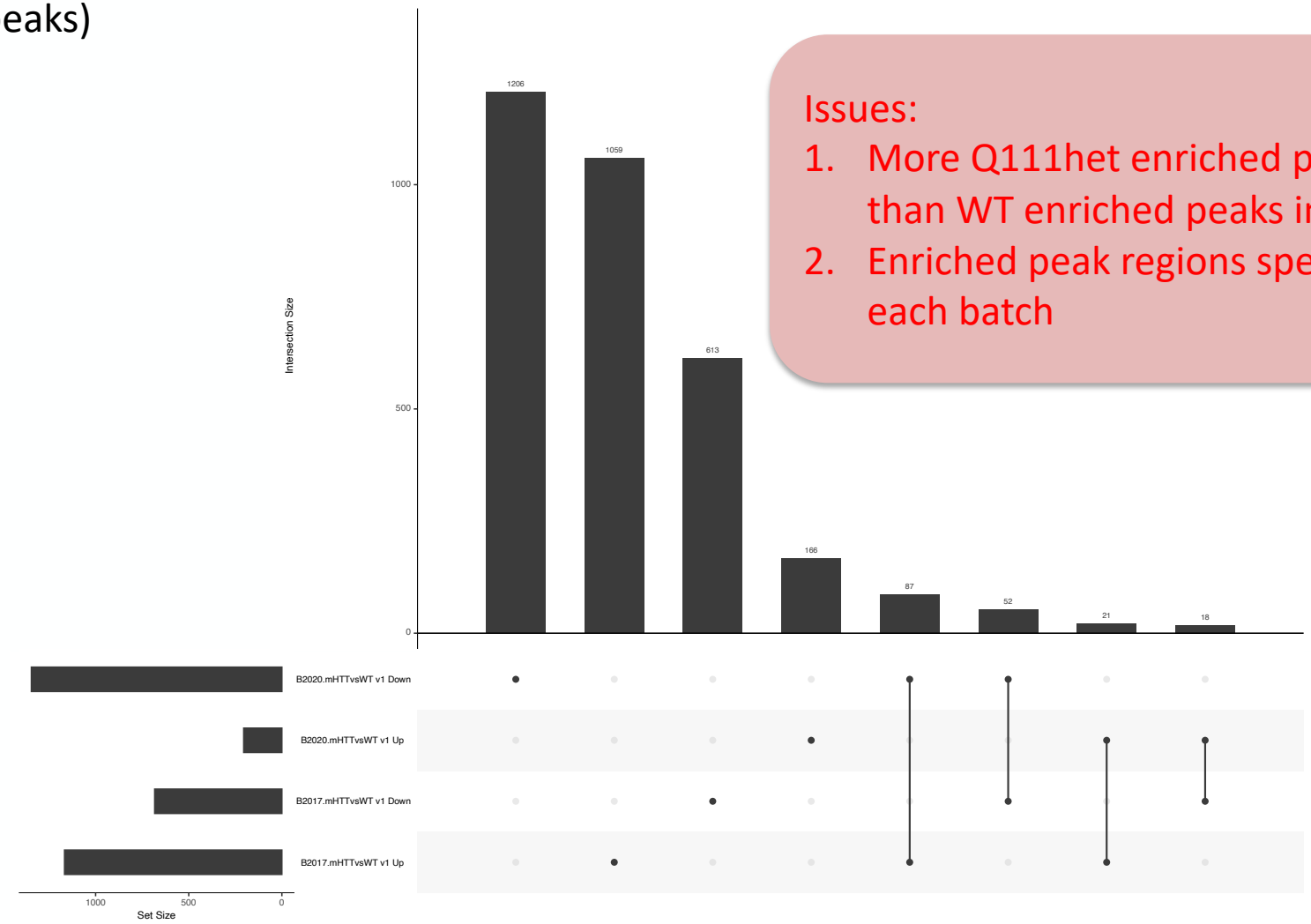
- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off



UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off (excluding LowReproducibility peaks)



Issues:

1. More Q111het enriched peaks than WT enriched peaks in B2017
2. Enriched peak regions specific to each batch

Version 3: p1e3 shared by ≥3 samples
(latest reproducibility groups)

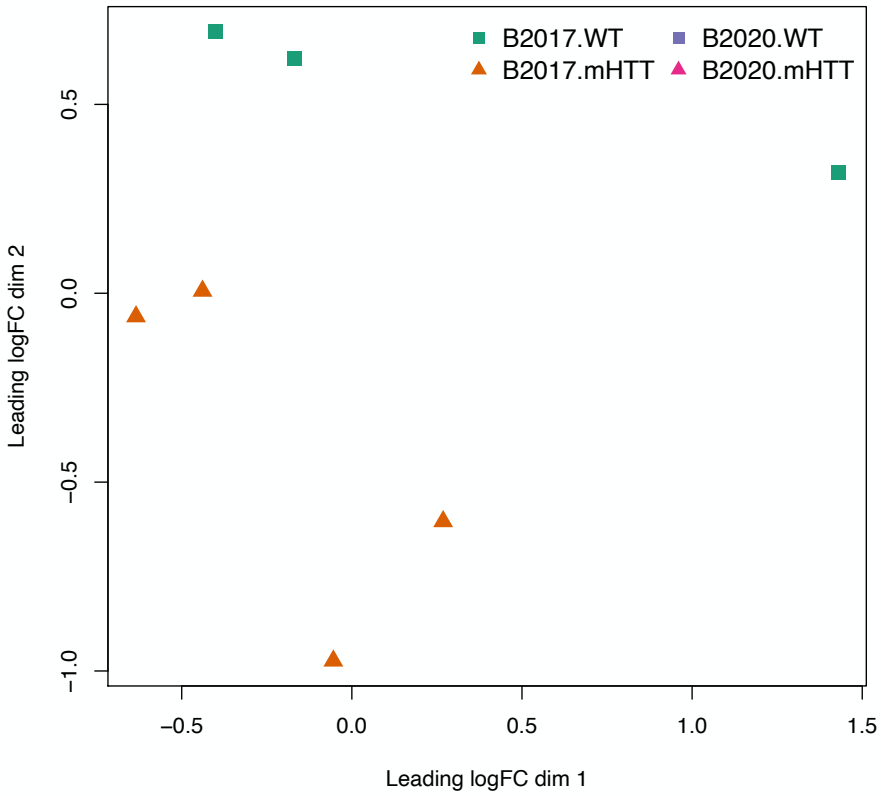
BATCHES ANALYZED SEPARATELY

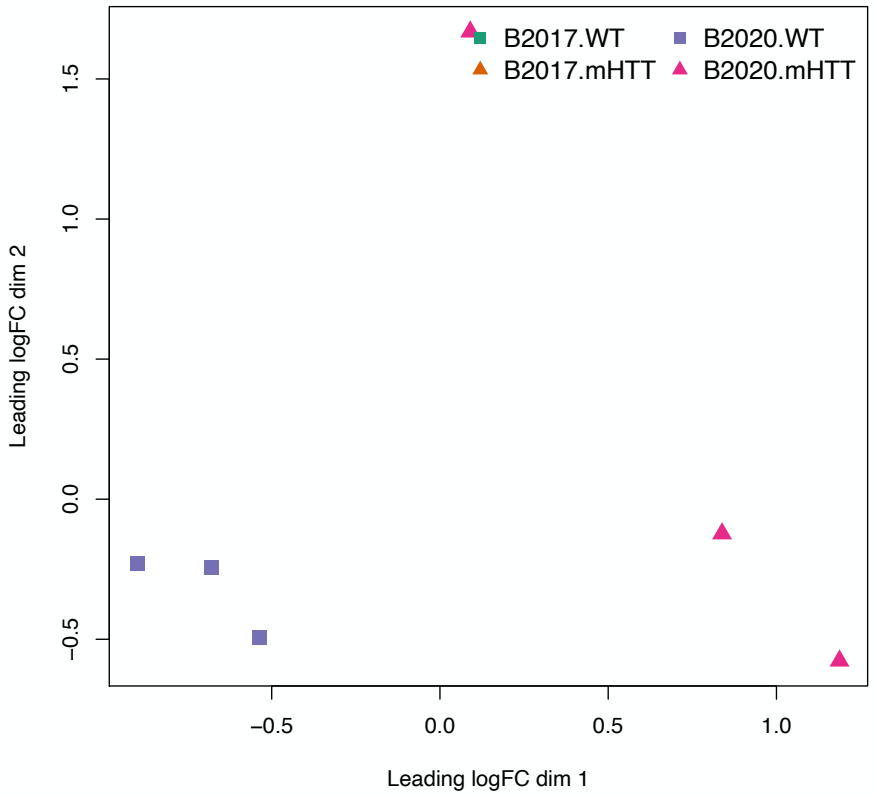
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Sample clustering (B2017 and B2020 analyzed together)

B2017 samples	B2020 samples





UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Differential binding between genotypes

Comparison	Significance Cut-off	# Regions LFC > 0	# Regions LFC < 0
2017 mHTT vs 2017 WT	P-value < 0.05	193	361
2017 mHTT vs 2017 WT	FDR < 0.05	0	0
2020 mHTT vs 2020 WT	P-value < 0.05	424	2,476
2020 mHTT vs 2020 WT	FDR < 0.05	1	0

Better: More WT enriched peaks than mHTT enriched peaks in both batches

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → peak counts from ~30M downsampled reads
- Differential binding between genotypes

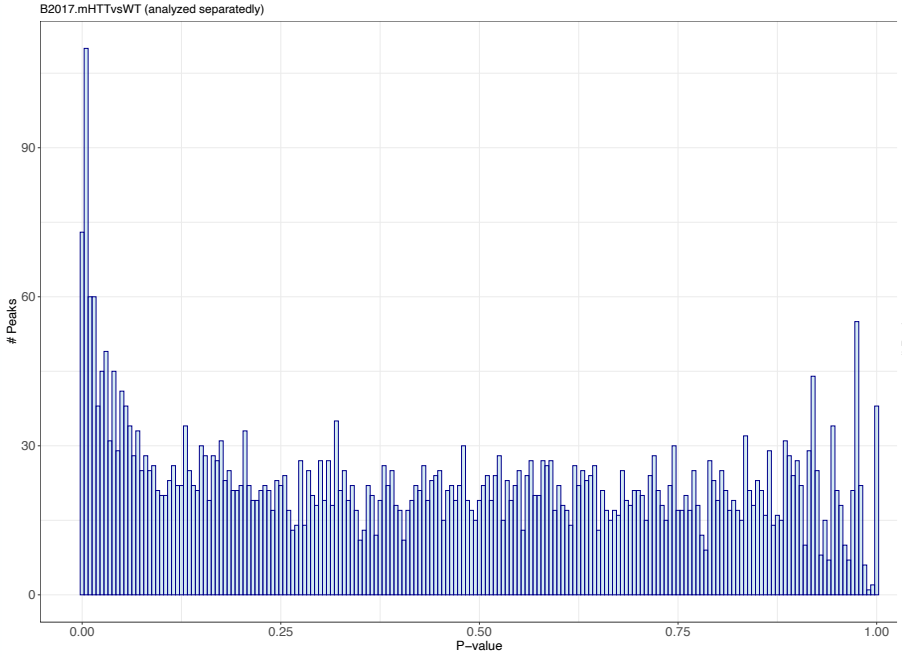
Htt_ChIPSeq.downsample_30000000.samples.diffbind.v20210831.separate_batches.HTT.diffbind_results.pdf

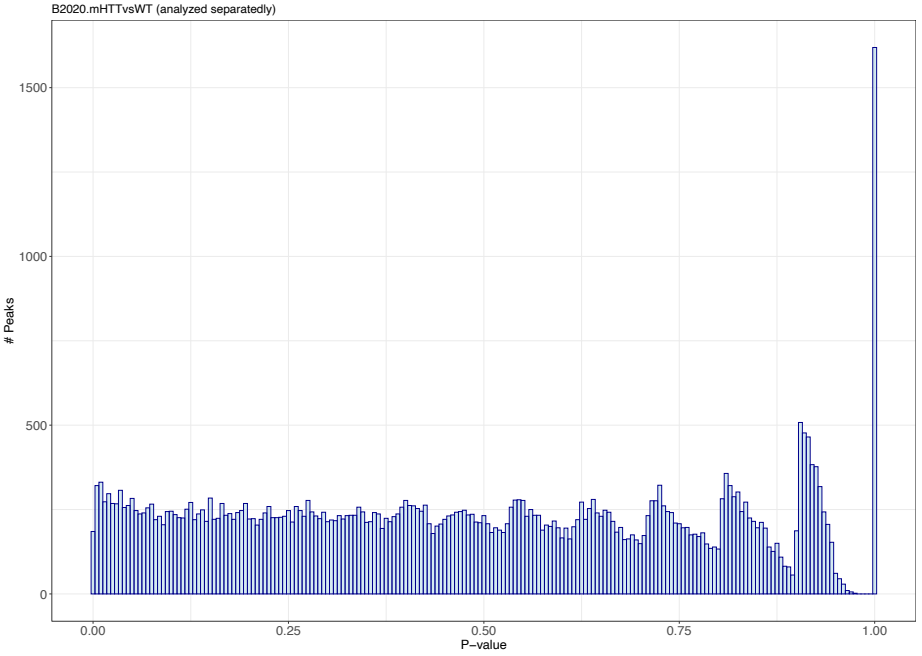
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value for each Peak Region

B2017 mHTT vs WT	B2020 mHTT vs WT



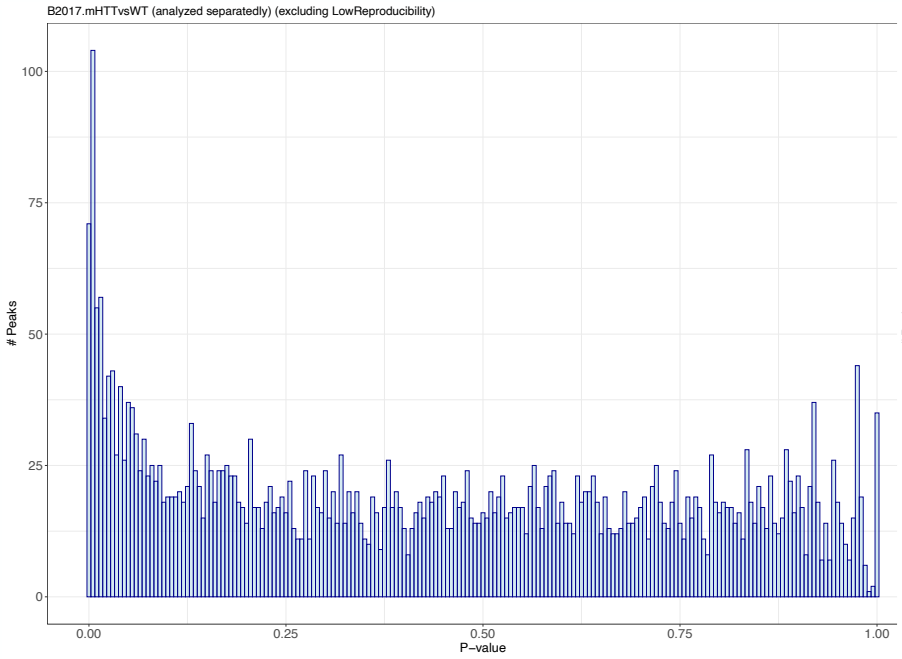


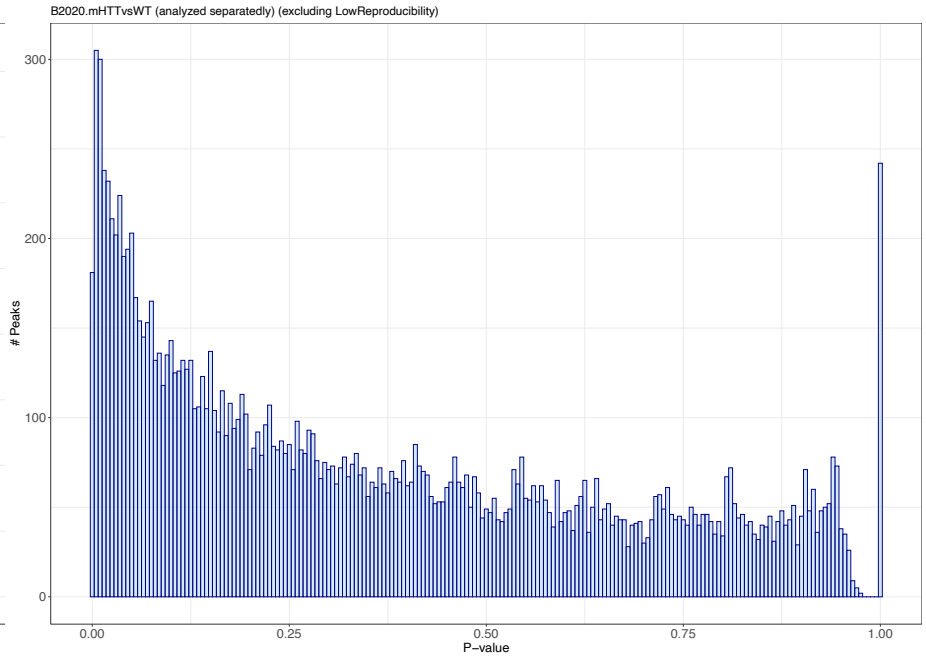
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value for each Peak Region (excluding LowReproducibility peaks)

B2017 mHTT vs WT	B2020 mHTT vs WT

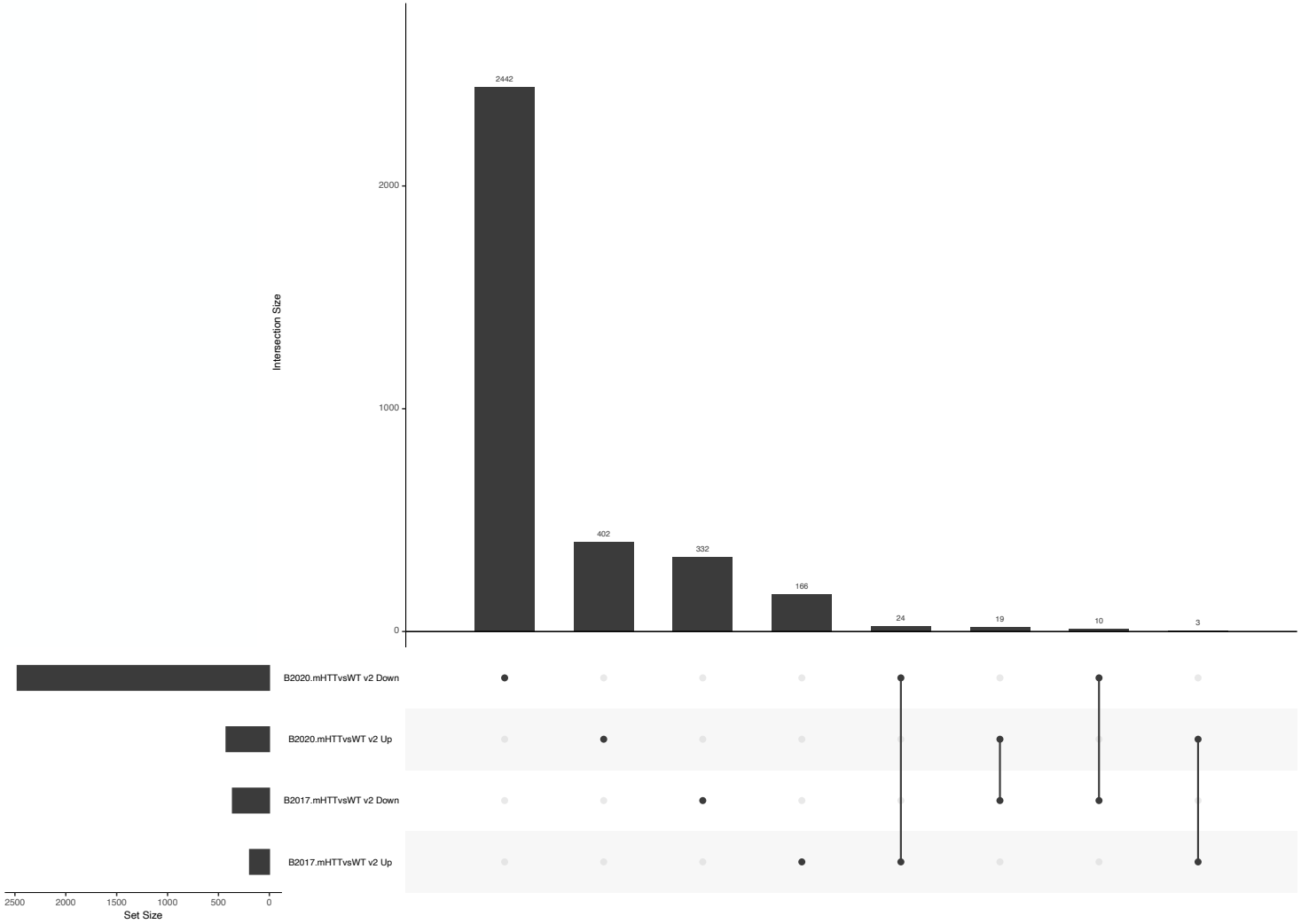




UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off

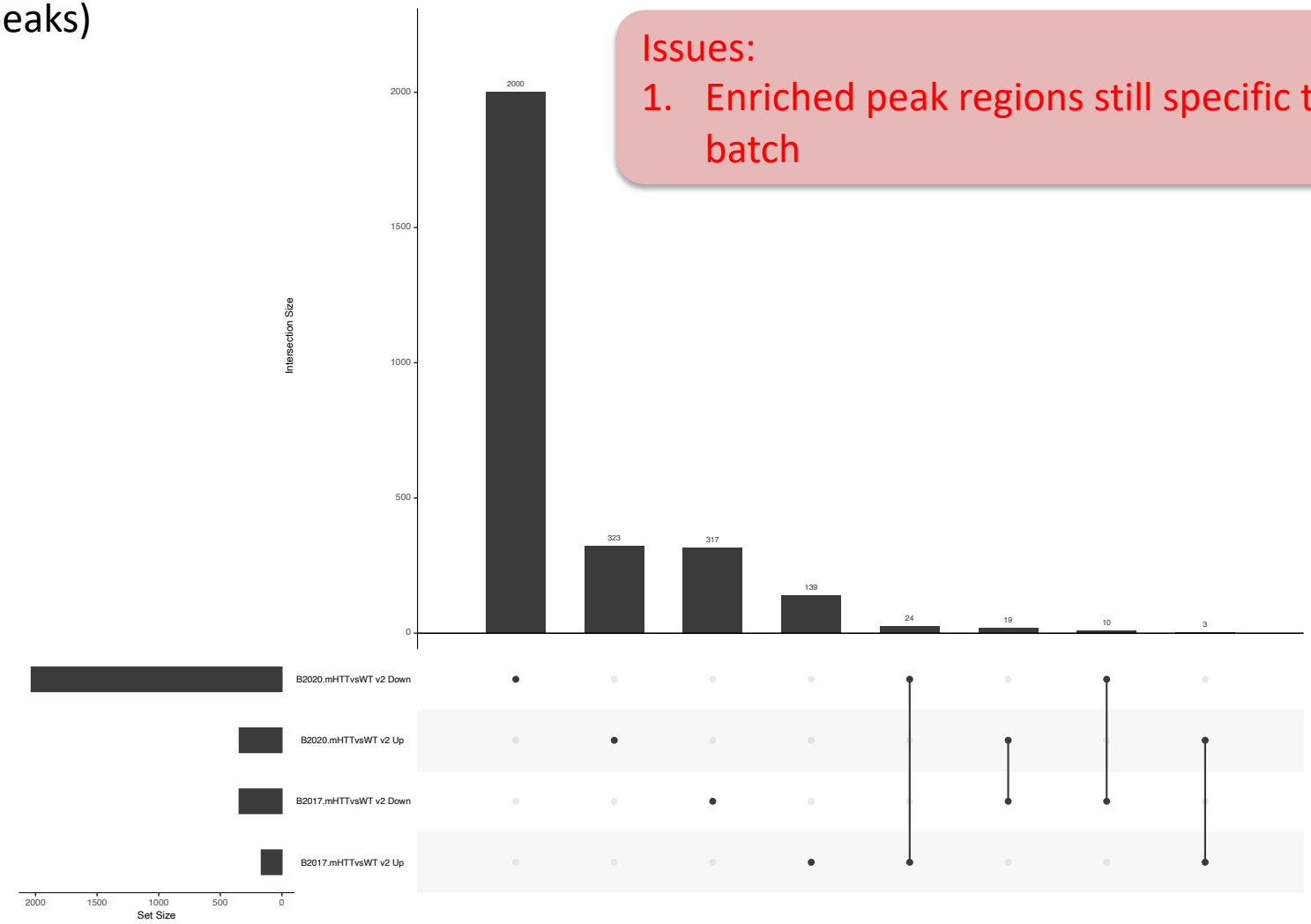


UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off (excluding LowReproducibility peaks)

Issues:
1. Enriched peak regions still specific to each batch



Version 3: p1e3 shared by ≥3 samples
(latest reproducibility groups)

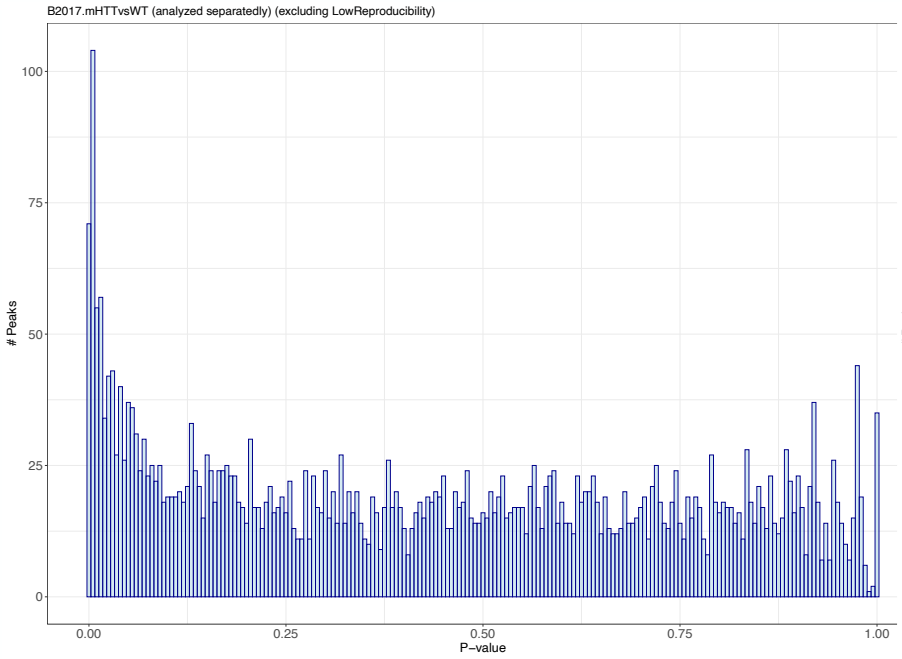
BATCHES ANALYZED SEPARATELY → META ANALYSIS

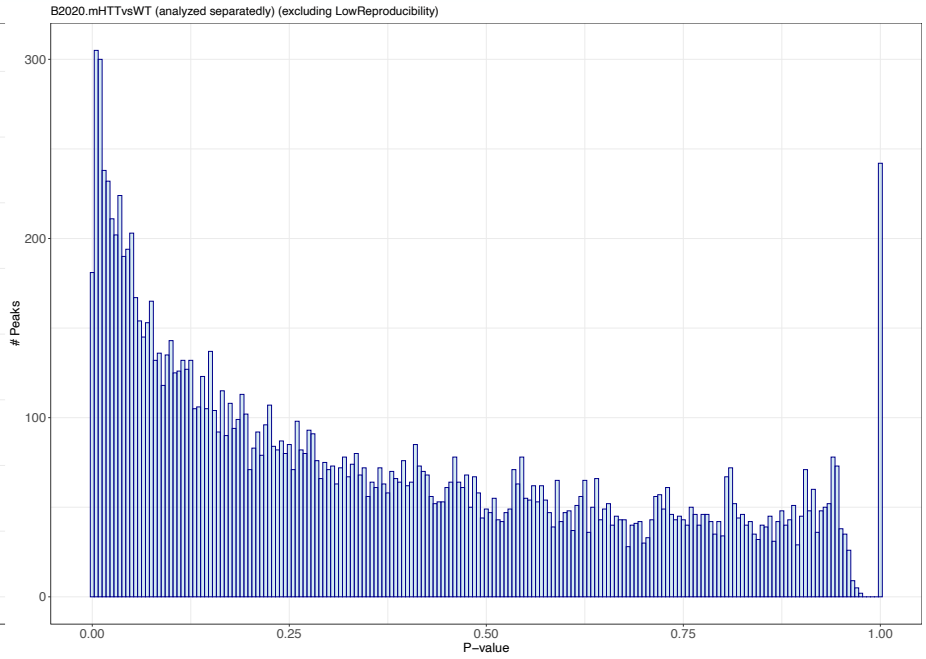
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Differential Binding

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value (from individual batches) for each Peak Region (excluding LowReproducibility peaks)

B2017 mHTT vs WT	B2020 mHTT vs WT



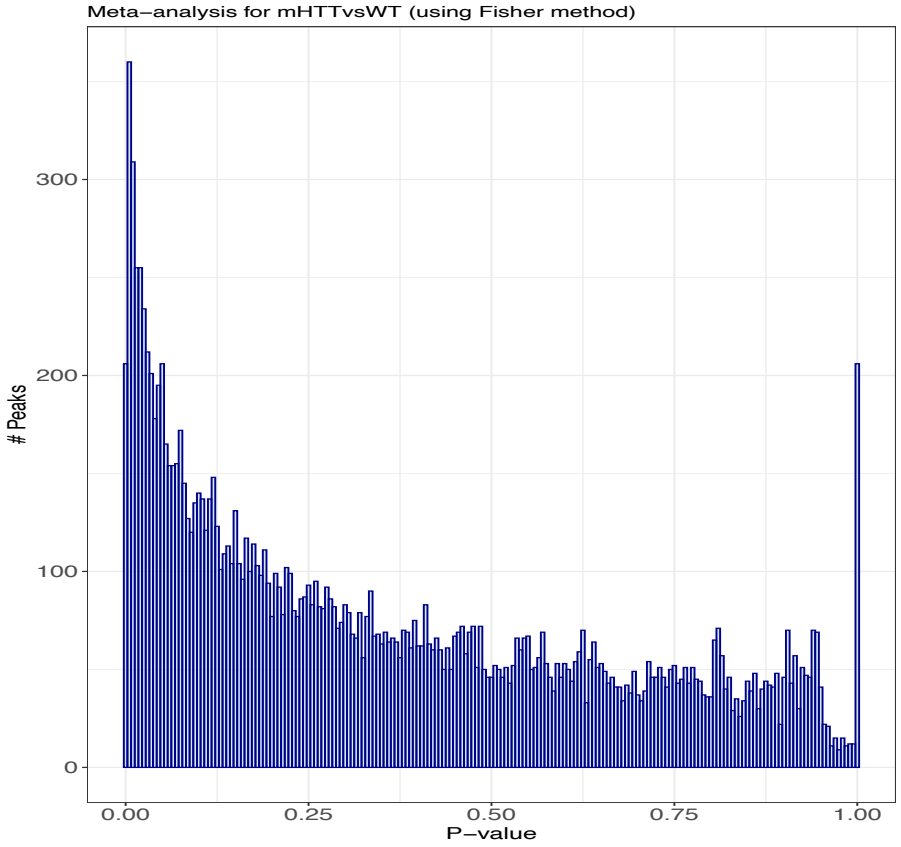


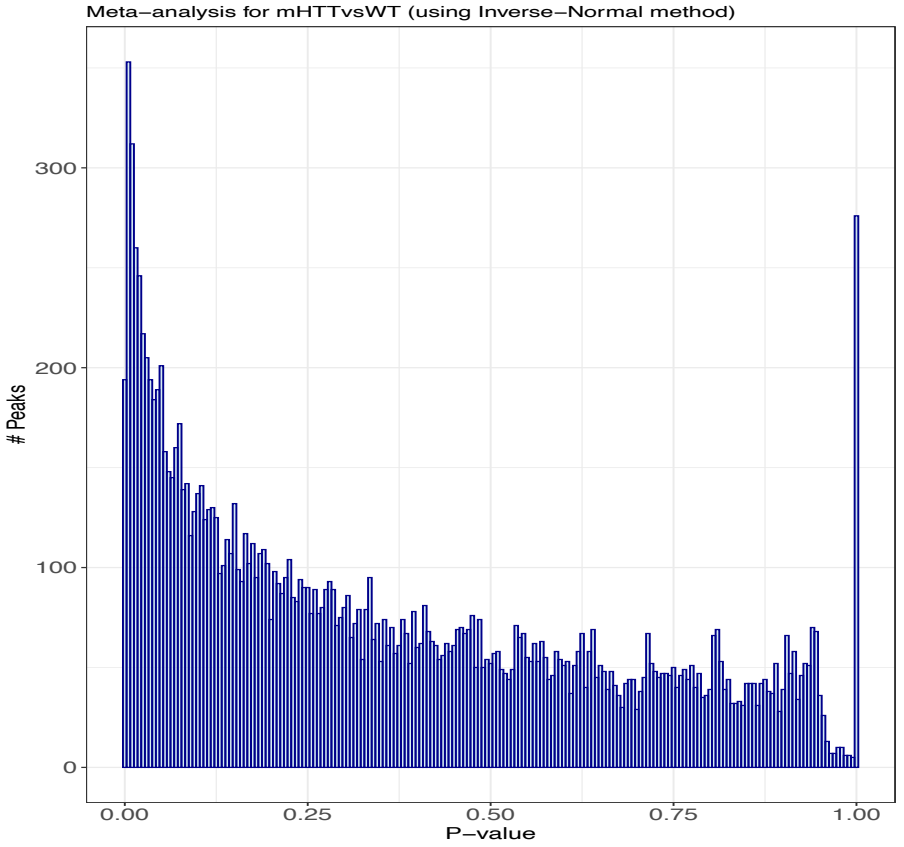
UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

DiffBind Meta Analysis

- 57,787 total peak regions → Differential binding between genotypes
- Distribution of P-value (from meta analysis) for each Peak Region (excluding LowReproducibility peaks)

B2017 mHTT vs WT	B2020 mHTT vs WT

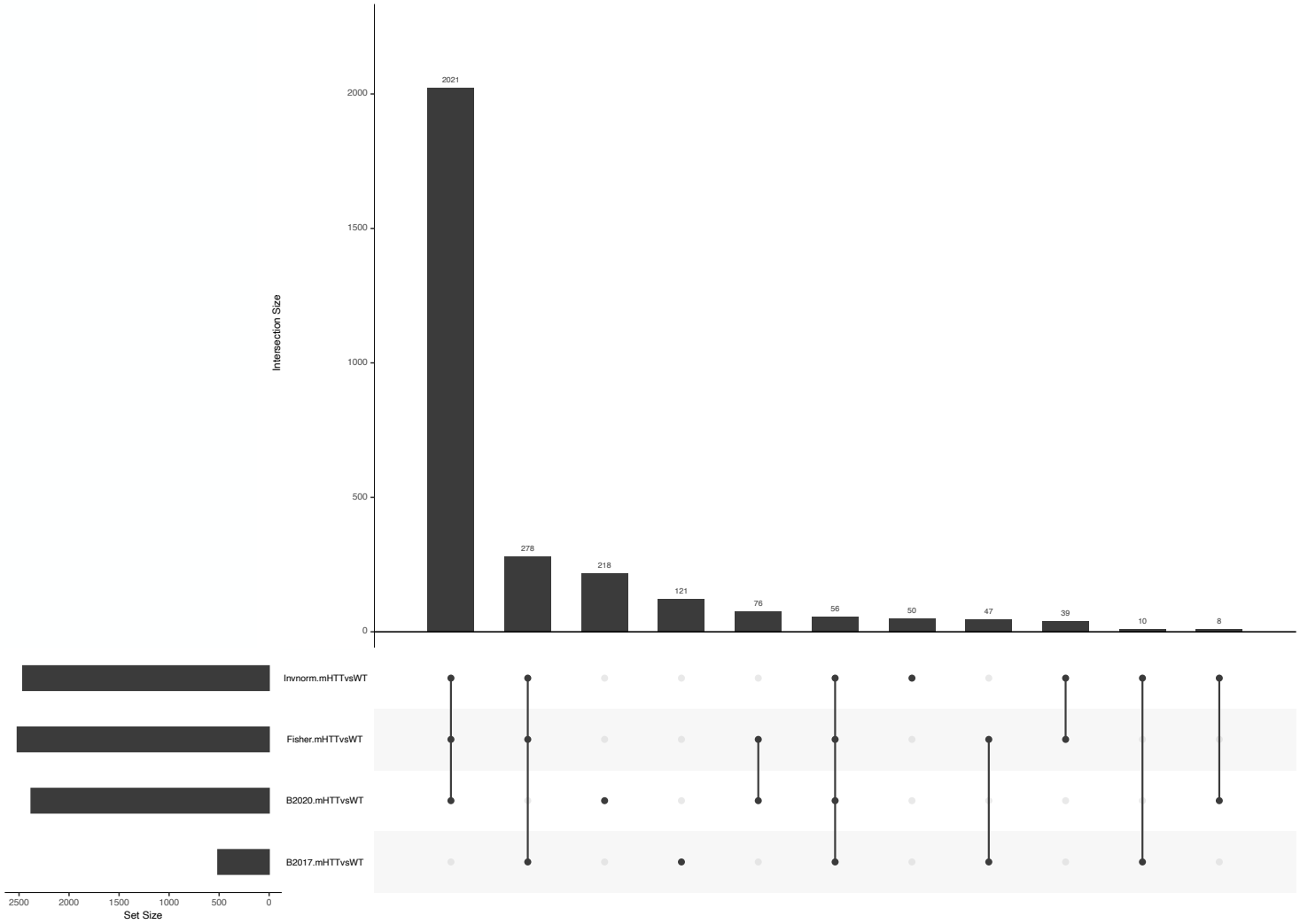




UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

DiffBind Meta Analysis

- 57,787 total peak regions → Differential binding between genotypes
- Overlap before and after meta analyses using P-value < 0.05 cut-off

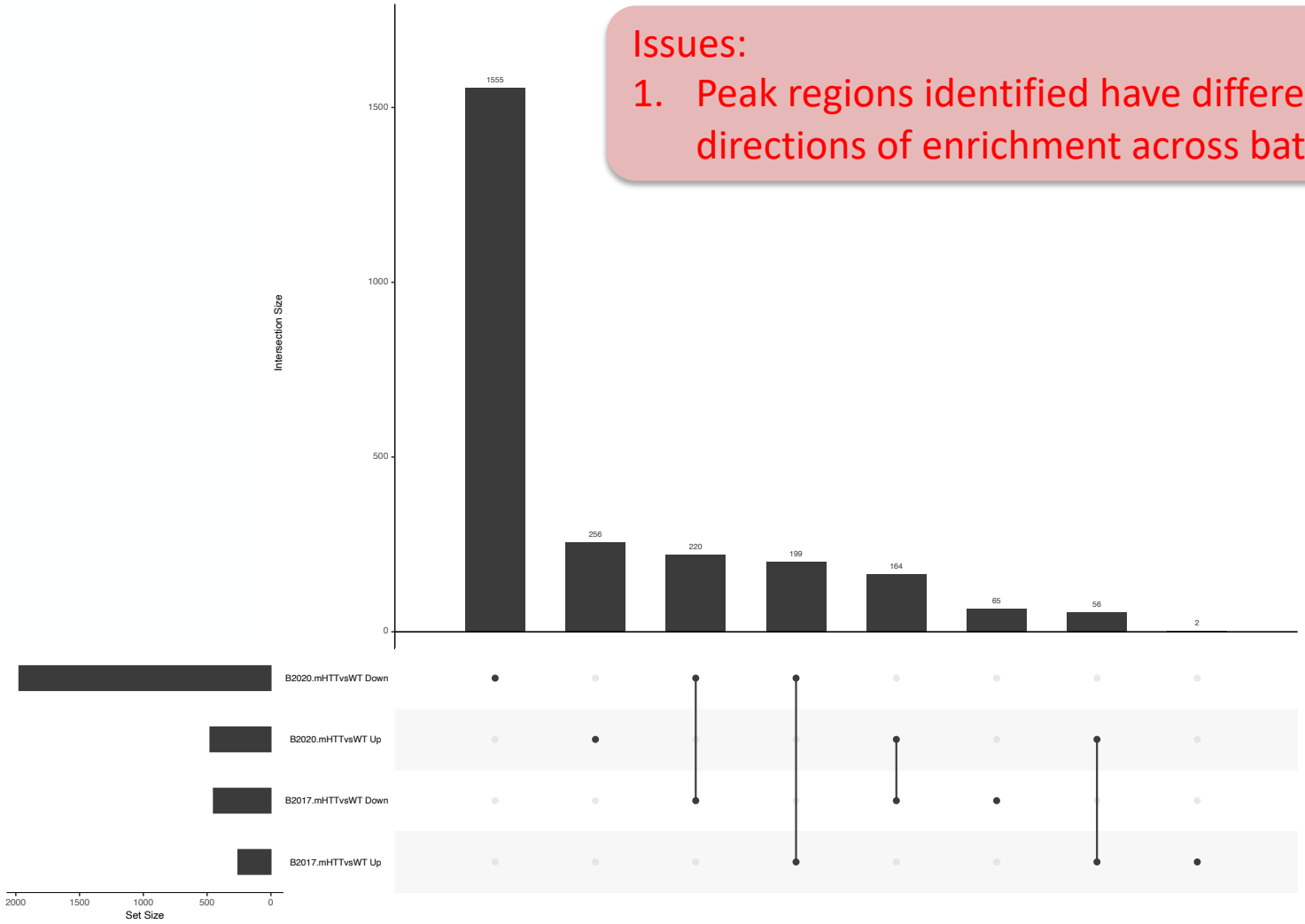


UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

DiffBind Meta Analysis

- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off (from Fisher method)

Issues:
1. Peak regions identified have different directions of enrichment across batches

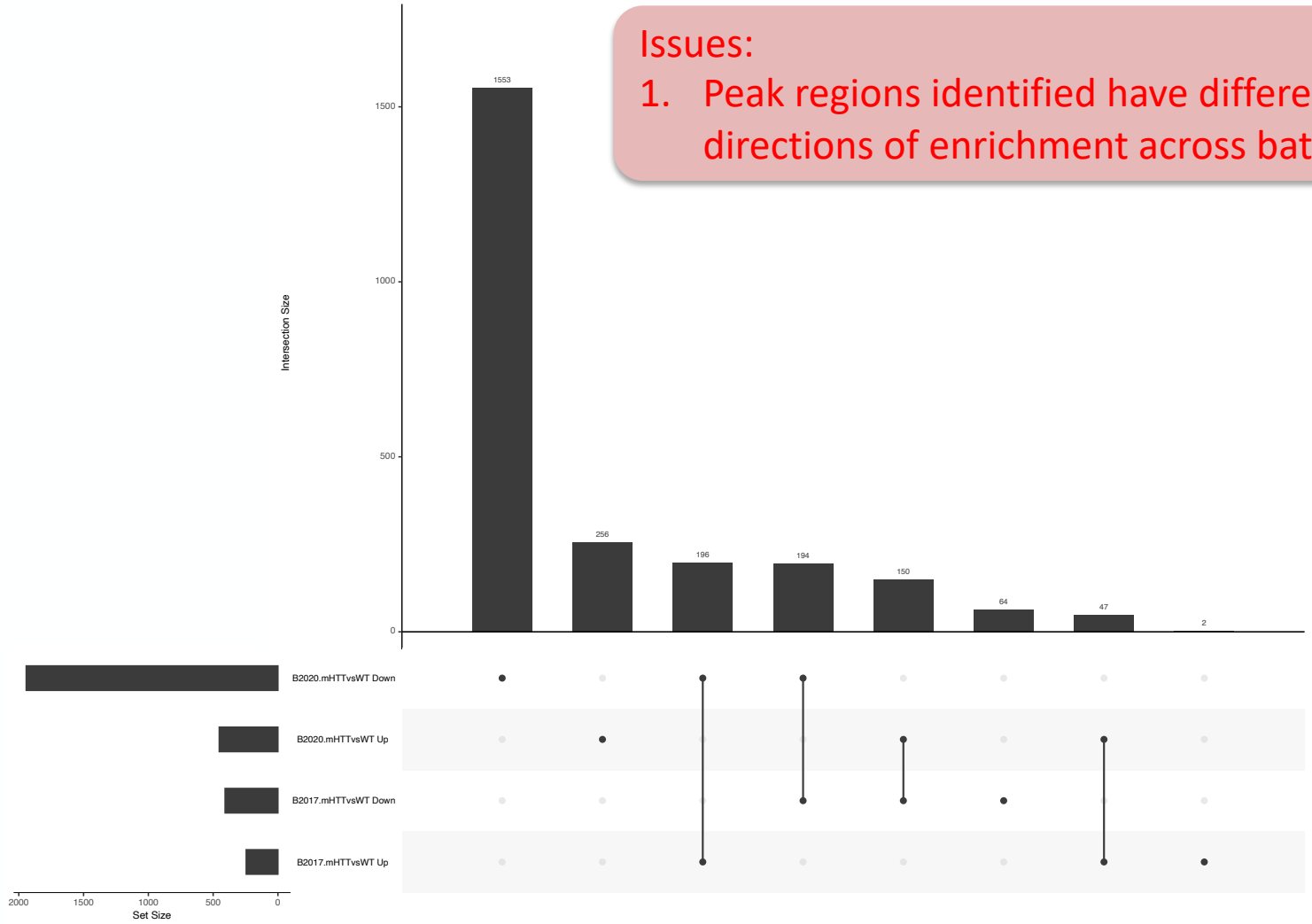


UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

DiffBind Meta Analysis

- 57,787 total peak regions → Differential binding between genotypes
- Overlap across batches using P-value < 0.05 cut-off (from Inverse Normal method)

Issues:
1. Peak regions identified have different directions of enrichment across batches



FUTURE STEPS

UNIVERSITY of MARYLAND
SCHOOL OF MEDICINE

Downstream Analysis

- On-going analysis (Questions)
  - How do we know that binding regions from the 2017 Q111het samples represent mHTT and not WT?
  - If 2017 Q111het samples is enriched for WT compared to mHTT, then wouldn't we see less overlap between B2017 and B2020 differential binding?
  - What peak sets should we use for downstream enrichment analyses?
- Future analysis (based on relevant set of peak calls)
  - Motif Analysis
  - Enrichment against different reference sets (ChromHMM, HDSigDB, etc)
  - Enrichment of Histone Mark regions
    - Existing ActiveMotif Histone Mark regions
    - Additional data from ActiveMotif for H3K9me? Delivery date?
  - Note: GAT Analysis is resource intensive and time consuming
    - Hence heavily dependent on available grid resources and grid load
    - The more iterations I run, the slower the analyses will get with each iteration due to lower grid priorities

QUESTIONS?