Zang Lab Research Project

William Franz Lamberti [1]

University of Virginia

Spring 2021

[1] MS Statistical Science
PhD Computational Sciences and Informatics

Outline

Introduction

- Goal: Explore transcriptional regulatory function of androgen receptor (AR) since it's important for prostate cancer
- Solution: Developed a pipeline for the analysis (Question 5, option 2)
- Still learning a lot
  - Background is Statistics and Images
  - Please provide constructive criticism so I can improve!

ChIP-seq: Software and Parameters

- `bowtie2_Build_h38_index.slurm`
  - `bowtie2-build`: built reference genome
- `fastq_to_bam.slurm` via `chipseq_analysis_on_input_file.sh`
  - `fastqc`: initial quality control (QC)
  - `Bowtie`: creating SAM files
  - `samtools`: converting SAM to BAM, sorting, filtering duplicates, and creating indices
- `macs.slurm`
  - `macs`: callpeak analysis
    - q (minimum FDR cutoff): 0.01
    - g (genome size): hs $= 2.7 \times 10^9$
- `r_chipqc.r`
  - `ChIPQC`: Produces plots and values for QC
  - This part was done locally, but code is ready for deployment

ChIP-seq: Software and Parameters References

- Harvard Chan Bioinformatics Core Workshop:
  - `https://hbctraining.github.io/Intro-to-ChIPseq/lessons/04_automation.html`
  - `https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html`
  - `https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html`
- Mostly good, but more guidance is needed for `ChIPQC` `R` package setup as it apparently only works with R Versions $<$ 3.6
  - Rivanna Tech reportedly got this to work for newer versions of R

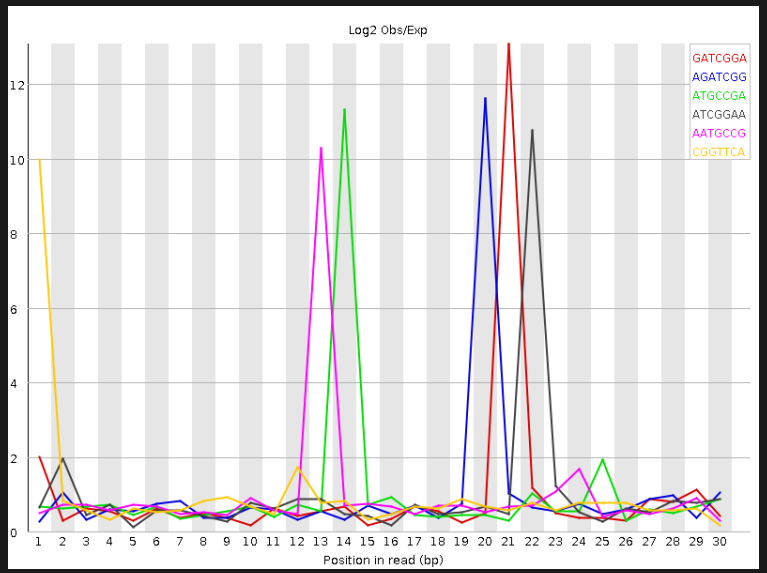
ChIP-seq: Quality Control (QC) Measures

ID	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%
Rep1	2,944,159	0	36	162	9.4	0.52	2.3
Rep2	2,527,581	0	36	160	7.9	0.5	2.4

- High Level: Each replicate produced very similar values
- Reads in Peaks (Peak Overlap Measure)
  - Transcription factor considered good if $> 5\%$
  - This is about half of that rule-of-thumb
- Duplicate Rate (Dup%)
  - If binding sites, occur at same location, will have large Dep%
  - We filtered these out during the analysis, which is why ours is 0%
  - Removing duplicates important when these is a small amount of starting material

ChIP-seq: `fastqc` Summary

- All plots produced mostly the same looking plots
- Per base sequence quality for DHT 2 was concerning
- All had concerning Kmer content plots
  - Shows if certain sequences occur too frequent
  - Should be all low values
  - This has some spikes in certain locations
- TLDR - Fairly confident that they are good samples



RNA-seq: Software and Parameters

- `rnaseq_complete.slurm`
  - `hisat_setup.sh`
    - `python 3.6.8`
    - `hisat2-build`
  - `rna_hisat2.sh`
    - `fastqc`
    - `hisat2`: Mapping each sample
    - `samtools`: Convert SAM, sorting BAM
  - `rna_stringtie_merge1.sh`
    - `stringtie`: assembling
  - `rna_stringtie_merge2.sh`
    - `mergelist_maker.sh`
    - `stringtie --merge`
    - `gffcompare`
  - `rna_stringtie_part3.sh`
    - `stringtie`: output files for `ballgown`

RNA-seq: Software and Parameters

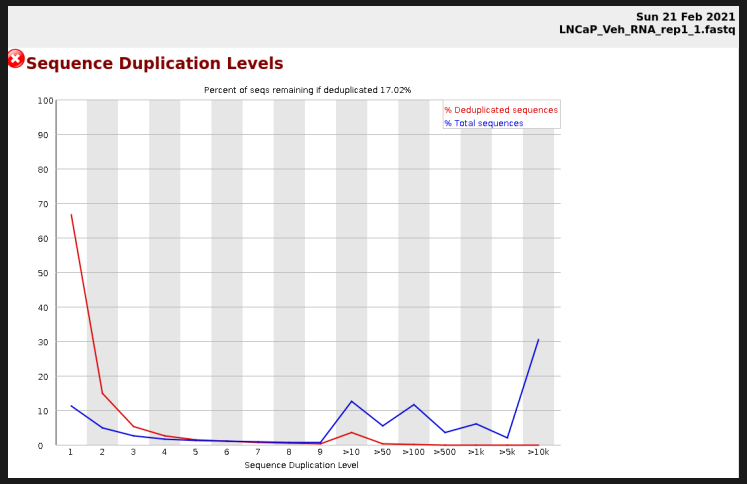
- `rna_ballgown.slurm`
  - `ballgown_csv_maker.sh`
  - `rna_ballgown.r`
    - Creates gene expression matrix
    - Differential Expression based on
    - FPKM
      - Used FPKM over RPKM
      - RPKM - single end
      - FPKM - paired end
      - `https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/`
    - Fold Change (FC)
    - $p-$value < 0.05

RNA-seq: Software and Parameters References

- Stringtie and Ballgown Paper
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5032908/
  - Example from blog: https://davetang.org/muse/2017/10/25/getting-started-hisat-stringtie-ballgown/
  - Advice: Do not save files in multiple different locations - put them all in one *MESSY* folder!
- Additional R code for plots
  - https://rstudio-pubs-static.s3.amazonaws.com/289617_cb95459057764fdfb4c42b53c69c6d3f.html
  - Provides a good baseline and example

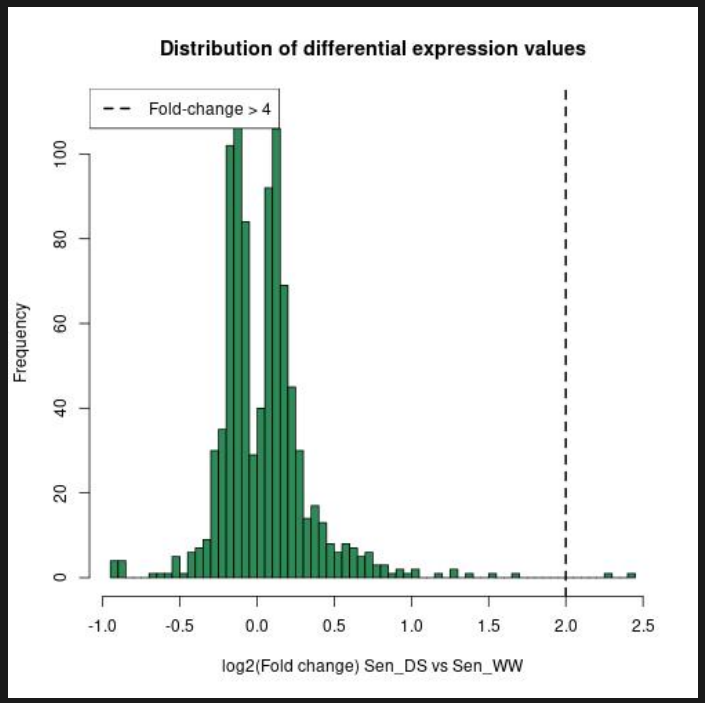
RNA-seq: `fastqc` Summary

- All plots produced mostly the same looking plots
- Had many different concerning plots
  - Kmer content plots
  - Sequence duplication levels
  - Per sequence GC content
  - etc.
- TLDR - Not confident that these are good samples



Joint Analysis: Sub-Q1 - Differential Expression (DE)

- Fold Change (FC)
  - FC measures difference between 2 quantities
  - FC > 4
- p-value < 0.05
- Note: x-axis title has an error in the labeling (should be log(FC) DHT vs Veh))



Gene Name	ID	FC	p-val	q-val
TMPRSS2	MSTRG.8523	5.4402	0.0029	0.3725
NKX3-1	MSTRG.12312	4.8245	0.0004	0.3725

Joint Analysis: Sub-Q2 - Genomic Distribution

Type	Proportion
N/A	> 0.00
3 UTR	0.02
5 UTR	> 0.00
Exon	0.01
Intergenic	0.42
Introns	0.50
non-coding	0.01
Promoter-TSS	0.02
TTS	0.02

Table 1: AR-1 and AR-2

Joint Analysis: Sub-Q3 - Promoter and Enhancer Sites

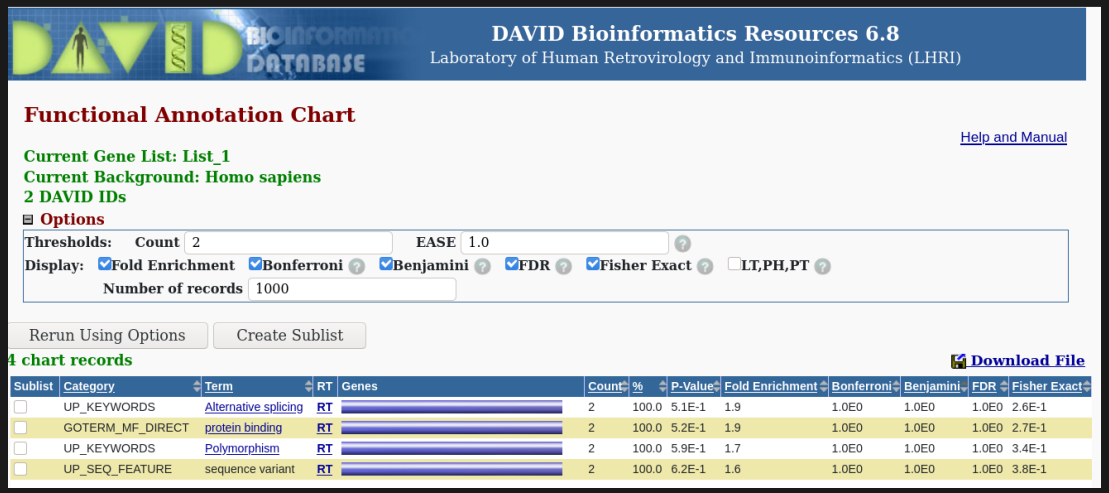
Type	Proportion
Down/Enhancer	0.41
Up/Promoter	0.59

Table 2: AR-1 and AR-2

Joint Analysis: Sub-Q4 - Comparing Sites for AR-1 and AR-2

- TMPRSS2
- Distribution
  - Intergenic: 1.00
- Up/Down
  - Down: 1.00
  - Up: 0.00
- NKX3-1
- Distribution
  - Intergenic: 2/3
  - Non-Coding: 1/3
- Up/Down
  - Down: 0.00
  - Up: 1.00

Joint Analysis: Sub-Q5 - DAVID



DAVID Bioinformatics Resources 6.8
Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Functional Annotation Chart

Help and Manual

Current Gene List: List_1
Current Background: Homo sapiens
2 DAVID IDs

Options
Thresholds: Count 2 EASE 1.0
Display: Fold Enrichment Bonferroni Benjamini FDR Fisher Exact LT,PH,PT
Number of records 1000

Rerun Using Options | Create Sublist

4 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Fold Enrichment	Bonferroni	Benjamini	FDR	Fisher Exact
	UP_KEYWORDS	Alternative splicing	RT		2	100.0	5.1E-1	1.9	1.0E0	1.0E0	1.0E0	2.6E-1
	GOTERM_MF_DIRECT	protein binding	RT		2	100.0	5.2E-1	1.9	1.0E0	1.0E0	1.0E0	2.7E-1
	UP_KEYWORDS	Polymorphism	RT		2	100.0	5.9E-1	1.7	1.0E0	1.0E0	1.0E0	3.4E-1
	UP_SEQ_FEATURE	sequence variant	RT		2	100.0	6.2E-1	1.6	1.0E0	1.0E0	1.0E0	3.8E-1

- Alternative splicing: Protein for which at least two isoforms exist due to distinct pre-mRNA splicing events
- Protein binding: Interacting selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other nonprotein molecules).
- Polymorphism: more than one allele (variation of same gene) occupies that gene's locus within a population
- Sequence Variant: Sometimes called mutation or polymorphism

Joint Analysis: Sub-Q6 - Motif Analysis via Homer

Homer *de novo* Motif Results (/scratch/tzp6pz /AR_1_peaks/)

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into ST
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 7321
Total background sequences = 40813
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-1137	-2.619e+03	37.66%	7.61%	42.0bp (67.1bp)
2		1e-1047	-2.413e+03	54.08%	17.99%	48.8bp (66.3bp)
3		1e-201	-4.629e+02	23.78%	11.20%	54.3bp (61.7bp)

Homer *de novo* Motif Results (/scratch/tzp6pz/ /AR_2_peaks/)

Known Motif Enrichment Results
Gene Ontology Enrichment Results
If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into STA
More information on motif finding results: HOMER | Description of Results | Tips
Total target sequences = 7320
Total background sequences = 40751
* - possible false positive

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background	STD(Bg STD)
1		1e-1067	-2.458e+03	38.31%	8.46%	42.3bp (65.9bp)
2		1e-957	-2.204e+03	63.95%	26.81%	52.0bp (66.2bp)
3		1e-236	-5.438e+02	73.55%	54.93%	54.1bp (65.5bp)

Conclusion

- Pipeline made for each RNA-seq and CHiP-seq
- Combined results to perform joint analysis
- Github link: `https://github.com/billyl320/zang_rotation_project`
- Future Work
- Further improve pipeline
  - Combine RNA/CHiP
  - Assume less about each sample file name
- Improve plots
  - Bigger fonts
  - Clearer labels/acronyms

Acknowledgements

- Dr. Zhenjia Wang
- Dr. Gladys Andino Bautista

Any Questions?

Extra

ChIP-seq: Cross-Correlation Plot

- Two Peaks
- Fragment length
  - Highest correlation value
  - About 150
- Read length
  - "Phantom" peak
  - About 35
- Very similar for each replicate
- Evidence of similar quality samples