In prostate cancer, androgen receptor (AR) binding and androgen-responsive gene expression are defined by hormone-independent binding patterns of the pioneer factors FoxA1 and GATA2. Insufficient evidence of the mechanisms by which GATA2 con- tributes to this process precludes complete under- standing of a key determinant of tissue-specific AR activity. Our observations suggest that GATA2 facilitates androgen-responsive gene expression by three distinct modes of action. By occupying novel binding sites within the AR gene locus, GATA2 positively regulates AR expression before and after androgen stimulation. Additionally, GATA2 engages AR target gene enhancers prior to hormone stimulation, producing an active and accessible chromatin environment via recruitment of the histone acetyltransferase p300. Finally, GATA2 functions in establishing and/or sustaining basal locus looping by recruiting the Mediator subunit MED1 in the absence of androgen. These mechanisms may con- tribute to the generally positive role of GATA2 in defining AR genome-wide binding patterns that determine androgen-responsive gene expression profiles. We also find that GATA2 and FoxA1 exhibit both independent and codependent co-occupancy of AR target gene enhancers. Identifying these determinants of AR transcriptional activity may provide a foundation for the development of future prostate cancer therapeutics that target pioneer factor function.

ChIP-seq reads were aligned against the standard hg19 build of the human reference genome with Bowtie v1.0.0 (38), allowing at most two mismatches. Reads with more than one valid alignment were removed, leaving only uniquely mapped reads for further analysis. Peaks were called using MACS v1.4.2 (39) with P-value threshold 1.0e-8. Heatmap counts were generated using HOMER v4.3 (40) with bin size 50bp. The raw ChIP-seq data have been submitted to the Gene Expression Omnibus (GEO) repository under the accession number GSE52725.

* Download the SRA files from ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP033/SRP033312/
* Convert the SRA files to either FASTA or FASTQ files (depending on which would be better) could possibly use SRA toolkit or other online resources.
* Could possibly dump the SRA files into R (look into possible packages for R).
* Once we convert we can use bowtie to map and align the files to the human Genome reference (could use galaxy or might need to download bowtie separately, using version 1.0.0 to stay true to original paper).
* We then can call peaks using MACS, will have to find version 1.4.2 (could possibly use most current version) MACS is used to find local bias in the genome sequence, meaning we can locate spikes in the data.
* HOMER v4.3 can be used to create heatmaps, can also try to pump the data into R or possibly python to create other graphs.

Before delving too far into the analysis we should first clean up the data too make sure we aren’t getting the most accurate results possible. Galaxy has a lot of great tools for cleaning up files. We can also get a quality report from FASTQ files in galaxy using FastQC so that is a possibility for helping clean up the data (refer back to courser videos for more on quality control and clean up).

<https://www.researchgate.net/post/What_methods_programs_are_recommended_to_clean-up_a_genome_from_contaminating_reads>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/>

<https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/chip-seq-analysis/chip-seq-practical>

<http://biocluster.ucr.edu/~rkaundal/workshops/R_feb2016/ChIPseq/ChIPseq.html#chip-seq-technology>

<http://www.abcam.com/webinars/a-step-by-step-guide-to-chip-seq-data-analysis-webinar>

<https://academic.oup.com/nar/article/42/6/3607/2437930/Three-tiered-role-of-the-pioneer-factor-GATA2-in#86892514>

<https://www.ncbi.nlm.nih.gov/books/NBK158900/>

How to calculate offsets for single

Trim off adaptors on the files