Creating an Epigenomic Map of the Heart

Jonah Poczobutt
New York University
jp6422@nyu.edu

Giulio Duregon
New York University
gjd9961@nyu.edu

Joby George
New York University
jg6615@nyu.edu

Abstract

Developing an improved understanding of gene expression is a growing focus in the scientific community. Currently, gene expression is thought to be driven by an ensemble of genetic information and cis-regulatory elements (CREs) working together. Our research goal was to create an epigenomic map of the human heart, identifying a universe of CREs and those which are differentially accessible when comparing sex, tissue type (atrium vs ventricle) and sex when controlling for tissue type. Through our research, we identified 21 K differentially accessible regions when comparing male ventricle tissue to female ventricle tissue. This expands upon previous work, showing that controlling for tissue type when performing differential accessibility analysis detects a higher level of genetic variability.

1 Introduction

On average, the human body is estimated to contain 100 billion kilometers worth of DNA. This large mass of information is condensed into our bodies because of the chromosomal structure Annunziato (2008).

147 base pairs of DNA wrap approximately twice around a histones protein core, forming nucleosomes. Groups of nucleosomes tightly bind to various proteins to make chromatin, which aggregates into a larger unit, a chromosome. For the most part, chromatin is tightly coiled, making it inaccessible to cis-regulatory elements (CREs). CREs are short, non-coding, DNA sequences that can bind to regions of open chromatin and enhance or reduce the replication of the associated open chromatin region. Lee 2011

Therefore, open chromatin regions play a large part in gene expression, and have been associated with a number of diseases, including cancer. The scientific community has been invested in creating an epigenomic map, locating open chromatin regions, their associated CREs, and investigating how downstream phenotypic outcomes relate to variability in this process. Complicating this effort is the fact that each tissue and its resident cell types have different gene and CRE interactions.

This project seeks to expand upon previous research by Lee et al 2018 in creating an epigenomic map of open chromatin regions and their associated CREs in the human heart. The human heart was chosen for two primary factors: previous research implicating CRE variation in cardiac disease, and the high morbidity of heart diseases. To accomplish this, we leveraged The Encode Project, an open-source database for tissue samples and DNA sequencing. We created functional maps based on sex, heart tissue type (atrium or ventricle), and sex when controlling for tissue type.

We then performed a differential accessibility analysis, which identifies open chromatin regions in one group of samples at a specific location of the genome, that are not present in another sample group.

Lastly, we wanted to tailor this paper to the broader data science community, rather than writing the report exclusively for bio-informatics researchers.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Related Work

Researchers have been working to understand the genetic mechanisms of diseases for decades. Two bodies of related work that are important to review are the varying methodologies that have historically been employed to examine open chromatin regions, and differential expression.

2.1 Historical methodologies to identify open chromatin regions

There are three primary scientific methods to identify open chromatin regions: chromatin immunoprecipitation with deep sequencing (ChIP-Seq), deoxyribonuclease I sequencing (DNase-Seq), and assay for transposase-accessible chromatin sequencing (ATAC-seq) Zhang Z (2011).

Out of the three methods, ChIP-Seq was the first developed in 2007, and provides a highly accurate way to detect in open chromatin regions Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007). The drawback of using this analytical methodology is the need for specific antibodies in the immunoprecipitation process, which makes the technique less generalizable across different tissue and cell types.

DNAse-Seq was the second analytical technique developed in 2008. The technique treats the tissue sample with DNase I enzyme, and uses high-throughput sequencing on the sample. DNAse I binds with accessible DNA and digests it, highlighting where open chromatin regions exist. The open chromatin regions identified by this process are also referred to DNase I hypersensitive sites (DHS). Compared to ChIP-Seq, this methodology is more generalizable, but less precise Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE (2008). Given the research goal of mapping the human heart's CREs and open chromatin regions, this technique is well suited to our analysis.

Lastly, ATAC-Seq is the most recent technique, developed in 2013 Buenrostro, J., Giresi, P., Zaba, L. et al (2013). The technique uses bacterial Tn5 transposase preloaded with sequencing adapters to identify open chromatin regions when paired with sequencing. This technique is also appropriate for the research goal, but was not used in our analysis. When considering future research endeavors, using ATAC-seq would help cross-validate the results found in this paper.

2.2 Differential Expression

Differential expression is when the observed frequency of occurrence of some gene varies between conditions. In our case, gene expression is quantified by mapping RNA-seq data to a reference genome. Understanding gene expression is a vital goal in the scientific community as many diseases are influenced by gene expression. Further research could be completed to analyze how differentially accessible regions of open chromatin impact differentially expressed genes.

3 Problem Definition and Algorithm

3.1 Task

The research objective of this project was to build an epigenomic map of the human heart and identify areas that are differentially accessible when comparing sex, tissue type, or a combination of sex and tissue type. Figure 1 shows the composition of the ENCODE datasets by gender, and tissue type.

Gender	Tissue Type	
	Ventricle	Atrium
Male	6	1
Female	4	3

Table 1: Breakdown of the sample data by different conditions, gender and tissue type

The first step in finding differentially accessible areas was to read raw DNA sequence bam alignment files, containing nucleotide base pairs (adenine, cytosine, guanine, and thymine) for the 14 samples

and identify per sample open chromatin regions. We used hotspot2, an open-source software developed at the Altius Institute for Biomedical Sciences in Seattle, WA, to detect sites along a genome sequence for high levels of enrichment, which identifies open chromatin regions.

Differing data quality between samples, and individual variation in the location of open chromatin requires researchers to create a consensus map of the samples before determining differential accessibility. A consensus map is a method of aggregating individual samples varying open chromatin regions into a singular collection of possibly accessible regions, and represents the totality of open chromatin regions of all the samples. The Index software designed and implemented by Wouter Meuleman and Eric Rynes takes individual DHS files as input, and merges them together to create a consensus map. We show our consensus map, along with two samples raw DHS input in Figure 1.

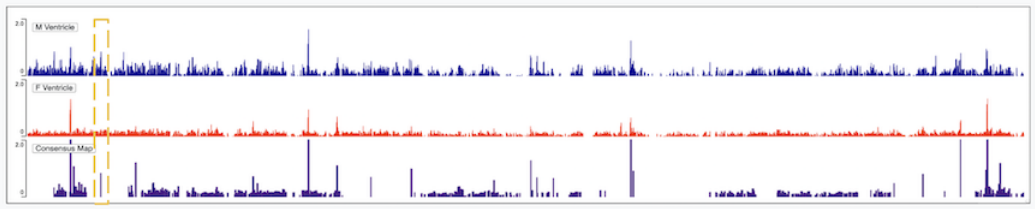

Figure 1: Visualization of the consensus map formed from all 14 samples, with two samples from the ventricle shown for visual clarity. The orange box highlights an area of differentially accessible chromatin between the male ventricles sample and female ventricle sample

Next, we read in the individual DNase-seq alignment files through the RSubread package, aligning specific open chromatin regions with the consensus map. This allows us to determine levels of expression across all possibly accessible regions for each sample. We obtain counts of how often these possibly accessible regions are observed to be accessible for each sample at our disposal.

Lastly, we performed a differential accessibility analysis, comparing the open chromatin regions for one condition (i.e. atrium) versus the open chromatin regions for the alternative condition (i.e. ventricle) using the DeSeq2 software.

3.2 Software & Algorithm

To accomplish our task, we used several popular bio-informatics software packages, outlined below, and visualized in Figure 2. We also wrote R and Python source code from scratch, available on our group's GitHub.

3.2.1 HotSpot2

In order to build a consensus map of the cardiac genome, (described further in Section 4.3), we needed to identify CRE hotspots and build a "consensus mapping" of their location around the genome. For CRE location discovery, we used hotspot2, an open-source software developed at the Altius Institute for Biomedical Sciences in Seattle, WA, used to detect sites along a genome sequence for high levels of enrichment, (DNA that has open chromatin region and can be bound to by mRNA). Hotspot2 is typically used to build maps of CREs along the genome on an donor id basis, who's output is potentially consumed or aggregated by a downstream process.

3.2.2 Index

Taking in the output of HotSpot2, we fed it into Index to create our consensus mappings. Index is a set of R scripts, designed and implemented by Wouter Meuleman and Eric Rynes, available in a public repository. The software takes individual DHS files as input, and merges them to create a consensus map of DHSs. This can be thought of as a universe of all potential DHSs, which may or may not be differentially expressed between conditions.

3.2.3 Rsubread

For our DNA and RNA sequencing tasks, we utilized the popular R software package, Rsubread Liao et al. (2019). Rsubread offers high performance DNA and RNA sequencing APIs accomplished

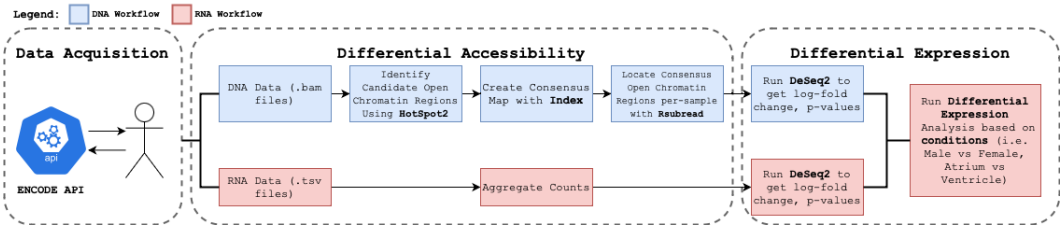

Figure 2: Project workflow visualized. Data acquisition is done through calls to ENCODE's public API, then fed through two separate tracks of preprocessing for differential accessibility and differential expression analysis.

by generating hash tables for indexing the reference genome in question. Hash table keys are 16bp sequences which map to corresponding chromosomal locations. The indexing process allows for easy read alignment, read summarizing tasks, and gene expression/ DHS accessibility data analysis. Rsubread also quantifies its uncertainty of its alignment within reads with a probability score, making it possible to distinguish between high and low quality reads. This quality of Rsubread proved invaluable, allowing us to drop low-quality read samples, and refine our data population to a smaller high-quality subset. Using Rsubread we were also able to create a matrix of observed DHS peaks per sample when compared to our concensus map created by Index, as was necessary to enable differential expression and differential accessibility analysis.

3.2.4 DeSeq2

For differential analysis between conditions, we leveraged DeSeq2, an R software package for differential analysis of gene count data. Rather than use maximum-likelihood based solutions, the software uses "shrinkage estimation for dispersion and log fold change" Love et al. (2014) in tandem with Empirical Bayes priors which produce more stable analysis, automatic control for the amount of shrinkage, and trustworthy heuristics for outlier detection. Our group's use case for DeSeq2 was primarily identifying differentially accessible and differentially expressed DHS between conditions, (eg. Male vs Female, Left Atrium vs Right Ventricle), and analyzing the results based on condition direction (log-fold change) and p-value (using $\alpha < 0.05$).

4 Experimental Evaluation

4.1 Data

4.1.1 ENCODE Dataset

We will be analyzing ENCODE's et and Dunham (2012) high-quality, high-throughput short read sequencing of adult cardiac tissue. Our data set is comprised both of Open Chromatin (DNA-seq) data and Gene Expression (RNA-seq) data. In total, our data set is comprised of 14 adult individuals, 7 female and 7 male. The ages of the participants ranged from 40-60, with the mean/median age both approximately 53. For each participant we have two files at our disposal; one DNA-seq file, averaging **6GB** per file, and one RNA-seq file, averaging **11MB** per file. Files were imported from ENCODE to NYU Langone's HPC cluster, Big Purple, by using the ENCODE REST API. Each file contains genomic tissue data sampled from different parts of the human heart, allowing us to compare data generated form samples in different regions (i.e. right atrium auricular region tissue vs. left cardiac atrium tissue). A descriptive table of the dataset can be found in Table 5, in the Appendix A.3. After filtering out low read-quality samples, we arrived at our finalized dataset, mentioned prior in Table 1.

4.2 GENCODE Genome Annotation File

For a list of comprehensive gene annotations, we leveraged GENCODE's Frankish et al. (2018) human chromosome genome annotation GTF reference. GENCODE is an open-source repository for gene annotations for human and mouse genome data. This was necessary during connecting differentially expressed DHSs to specific gene identification.

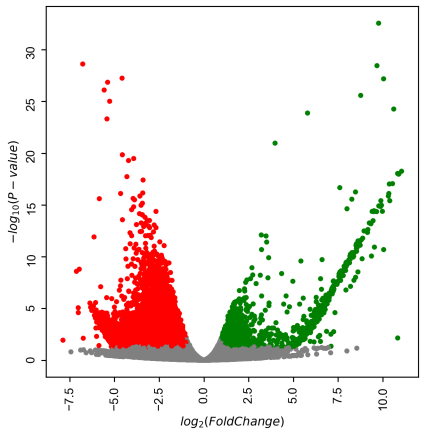

Figure 3: Volcano plot of differentially accessible DHS between male and female ventricle tissue. Red dots symbolize statistically significant differences in accessibility in female samples, while green dots symbolize the same but in male samples.

4.3 Methodology

We began our project by writing Python source code to query ENCODE's et and Dunham (2012) REST API to retrieve file metadata for pairs of DNA-seq, RNA-seq genomic data. After parsing our information, we could feed the API's response into another script that would retrieve the data files from ENCODE's website, and deposit them in our projects directory.

We then generated maps of DHS locations for each DNA-seq genome read we had available using HotSpot2 3.2.1. Taking the outputs, we fed them into Index 3.2.2, which produced our consensus maps. We removed "singletons" from our consensus map. These are consensus peaks that were only found in one of our samples. We did this to remove peaks likely resulting from random noise, and because these noisy peaks are much more likely to be differentially expressed (as they are only found in one sample, and therefore one condition.) This makes our results more robust to noise, at the potential expense of removing valid peaks that only happened to be found in one sample. After establishing our consensus map of DHS peaks, we leveraged Rsubread Liao et al. (2019).

Once the workflow with Rsubread was completed, we could run DeSeq2 Love et al. (2014), and quantify the statistical significance of the variance between conditions and genetic material.

We then filtered for statistically significant results, using a p-value of $\alpha = 0.05$, as based on the recommendation from our project advisors. We could then filter the data on log fold change, where positive and negative values would communicate which condition expressed a greater amount of variance for the genome sequence at any given point. In doing so, we created a subset of condition specific genomic variability, and could perform analysis on the RNA-seq reads for specific genes that would allow us to analyze if any causal relationship existed. The results are outlined in Section 4.4.

4.4 Results

We constructed out consensus map from all sexes and tissue types to identify around 336k non-singleton peaks. Removing singletons reduced the number of consensus peaks by more than half. We observe that as the number of samples contributing to the consensus map increases, the percentage of singleton peaks falls even as the total number of consensus peaks increases.

When examining ventricle data only and grouping samples based on sex, we identified 21,830 as being differentially accessible between the two conditions. Of these 14,862 were more highly accessible in females while 6,968 were more highly accessible in males. This can be seen visualized by the volcano plot in Figure 3.

Examining sex differences in a specific tissue type yielded more differentially accessible regions than comparing sex differences generally, or tissue type differences generally. This affirmed our intuition that subsetting our sex comparison to a particular tissue type would yield the most fruitful

Number of Samples	Total Peaks	Total non-singleton peaks	Percent Singletons
2	308496	77616	0.74
4	337255	131694	0.61
6	522047	214906	0.59
11	587317	277452	0.53
15	658144	328804	0.50
16	664301	336466	0.49

Table 2: Table of Singleton Peaks by Number of Samples

results, as we could remove noise resulting from certain DHSs having discrepancies in accessibility between tissue types. We identified only 13,672 differentially accessible DHSs between ventricle and atrium tissue, and only 6,426 differentially accessible DHSs when comparing sexes more generally. Essentially, noise from tissue type variability in accessibility can make it more difficult to identify differentially accessible regions between sexes. Restricting our comparison to a single tissue type (ventricle tissue), allows us to better capture sex differences. Further discussion of DHSs and further downstream analysis will be based on differentially expressed DHSs between sexes identified after subsetting to ventricle data.

Connecting differential accessibility to differential expression of certain genes was an important goal of our project. There are many potential comparisons that could be evaluated using the data we generated, but we restrict ourselves to two more straightforward questions here:

1. *Are DHSs more likely to appear in close proximity to gene regions within the genome?*
2. *Are more highly accessible DHSs more likely to appear in close proximity to upregulated genes than an identically sized region chosen at random from the same chromosome?*

The first question aims to determine whether DHSs (whether they are diferntially accessible or not) distribute uniformly within each chromosome or appear in close proximity to the genes they presumably help to regulate. To answer this question, we compared the count of DHSs within some neighborhood of a gene to the number of DHSs appearing in a random region of the same size, within the same chromosome. We collected the following results, showing that the mean number of DHSs within close proximity of a gene is greater than the mean number appearing in a random subset of the same size for all tested neighborhood sizes.

Condition	Neighborhood Size (kbp)		
	12.5	**25**	**50**
All Genes	55.44	108.84	215.36
All Genes R.S.	50.37	101.38	203.05

Table 3: Mean number of DHSs appearing within neighborhood of selected region

The second question performs a similar comparison using genes that are upregulated in a given condition, and peaks that are more accessible in that condition. We collected counts of more highly accessible DHSs appearing within a certain neighborhood of an upregulated gene, and compared these to counts of more accessible DHSs in a random region of the same size, within the same chromosome. Once again we observe a greater mean number of such DHSs in the neighborhoods of upregulated genes than we do in the randomly chosen region for all tested neighborhood sizes.

4.5 Discussion

One important result from this project is the enhanced granularity with which we can observe sex differences when we restrict our comparison to ventricle data only. Removing noise from variability in expression between tissue types. These results highlight the importance of following a similar procedure to the one we pursued that attempts to control for variability between tissue types. This result also seems to recommend coordinating data collection to ensure that common tissue types are extracted from many samples.

Condition	Neighborhood Size (kbp)		
	12.5	**25**	**50**
Male Up-regulated	1.58	3.1	5.97
Male Up-regulated	1.1	2.03	4.69
Female Up-regulated	2.32	4.63	9.01
Female Up-regulated	2.03	4.04	8.03

Table 4: Mean number of more accessible DHSs appearing within neighborhood of selected region

More samples to examine likely would have reduced the number of singleton peaks, thereby increasing the number of peaks in the consensus map. This could have helped make more concrete difference between upregulated genes and randomly chosen regions for example, however, the universe of differentially expressed genes would not change as a result of increasing samples from which we drew differential accessibility results.

Obtaining samples from individuals with specific negative health outcomes and comparing these to healthy individuals would also be interesting. We anticipate that a nearly identical process to the one we performed could we employed if this data were available.

Additionally, for future research efforts, we recommend replicating this analytical process using ATAQ-Seq data to confirm the importance of tissue type subsetting when performing differential accesibility analysis.

5 Conclusion

Better understanding the role of CREs in the human genome and discrepancies in the accessibility of these elements between groups is vital for better understanding genetic factors contributing for heart disease. Further work building on our results would likely incorporate known relationships between certain genes and their associated phenotypic relationships. One could also use our consensus map to correlate differential accessibility with certain health outcomes between individuals or groups, illuminating relationships between CREs and health outcome data.

6 Lessons learned

The biggest lesson in this research endeavor was the difficulty in cleaning the DNase-Seq data to get valid results. When trying to understand which open chromatin regions were differentially accessible, there was a long process in achieving scientifically valid results. The data quality of the DNase-Seq datasets lead to many spurious regions of open chromatin after the first attempt to perform the analysis. This lead to 660 K open chromatin regions being identified, which was far too high compared to previous studies and the expectation of our mentors. To remediate this issue of spurious open chromatin regions, we had to remove some biological samples, as they were low quality reads. Additionally, even for the higher quality samples, there were 330 K regions of open chromatin that were exclusively identified in a single sample (singletons). Per the advice of our mentors, we excluded them when running our differential accessibility analysis.

7 Student contributions

Jonah was in charge with all aspects of created the consensus map, running HotSpot2, Index, and Rsubread. Giulio performed the data scraping from Encode, and wrote source code to take the outputs of DeSeq2 and run experiments between conditions. Joby wrote the report, created the visualizations, and performed the differential expression analysis by running DeSeq2.

References

A Annunziato. 2008. DNA Packaging: Nucleosomes and Chromatin. *Nature Education*, 1(1):1–26.

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 129(4):823–837.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322.

Buenrostro, J., Giresi, P., Zaba, L. et al. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature Methods*, 10:1213–1218.

al et and I. Dunham. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. 2018. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1):D766–D773.

Yang Liao, Gordon K Smyth, and Wei Shi. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research*, 47(8):e47–e47.

M.I. Love, W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(550).

Pugh BF Zhang Z. 2011. High-resolution genome-wide mapping of the primary structure of chromatin. *Cell*, 144(2):175–186.

A Appendix

A.1 Software Dependency Download Links

1. R: https://www.r-project.org
2. hotspot2 https://github.com/Altius/hotspot2
3. Index: https://github.com/Altius/Index
4. Rsubread: https://bioconductor.org/packages/release/bioc/html/Rsubread.html
5. DeSeq2: https://bioconductor.org/packages/release/bioc/html/DESeq2.html

A.2 Project GitHub Link

Project GitHub: https://github.com/jp6422/dsga-capstone

A.3 Dataset Metadata Reference

Donor ID	Experiment DNA	DNA FileName	RNA Filename	Gender	Age
NCDO793LXB	ENCSR984SQJ	ENCFF554LAG	ENCFF132SDD	F	53
NCDO793LXB	ENCSR070CMW	ENCFF142DLY	ENCFF862LZL	F	53
NCDO271OUW	ENCSR278SKG	ENCFF805NZY	ENCFF940KYP	F	51
NCDO856ZOJ	ENCSR032NNU	ENCFF479NNN	ENCFF119FZL	F	59
NCDO856ZOJ	ENCSR395HAE	ENCFF445INC	ENCFF408FCD	F	59
NCDO856ZOJ	ENCSR622HTS	ENCFF116AFG	ENCFF840MWP	F	59
NCDO520EJG	ENCSR770OTB	ENCFF302BPE	ENCFF394BIS	M	60
NCDO520EJG	ENCSR895GSY	ENCFF580ZCH	ENCFF717WSV	M	60
NCDO520EJG	ENCSR085MZL	ENCFF720KNU	ENCSR015PUN	M	60
NCDO520EJG	ENCSR485UQY	ENCFF046LQL	ENCSR853TXT	M	60
ENCDO520EJG	ENCSR355WAJ	ENCFF870JXG	ENCFF651KGY	M	60
ENCDO411EVD	ENCSR747SEU	ENCFF126CZV	ENCFF440HGB	F	46
ENCDO411EVD	ENCSR374VQC	ENCFF155TZW	ENCFF434VRE	F	46
ENCDO411EVD	ENCSR356RNZ	ENCFF382FIL	ENCFF784GEM	F	46
ENCDO575WHY	ENCSR524QBS	ENCFF113NKE	ENCSR818DBU	F	41
ENCDO392CRK	ENCSR238FMP	ENCFF542RZV	ENCFF821VVG	M	40
ENCDO392CRK	ENCSR598RVJ	ENCFF350NTU	ENCFF311HGE	M	40

Table 5: Full Dataset MetaData, before filtering for low-quality reads