# Human Reference Genome Setup & Gene Feature Extraction (hg38): An OJT-Ready Pipeline for PRNP and NEFL

Author: Dr. Broderick Crawford

Date: August 15, 2025

## Abstract

This white paper outlines a reproducible, workstation-scale pipeline for constructing a human reference genome environment (hg38), integrating high-quality annotations (GENCODE v46), and generating transcriptome and coding sequence (CDS) datasets. The protocol further derives gene-focused assets for PRNP (prion protein) and NEFL (neurofilament light), including ±1 kb promoter windows and preliminary motif reconnaissance. Designed for on-the-job training (OJT), this methodology provides a practical, transparent workflow for neuroscience and rare-disease applications, operationalized on macOS using a Homebrew-managed bioinformatics stack.

## Introduction

Establishing a robust local reference genome environment is a critical prerequisite for modern genomics research. The human reference genome (hg38) remains the most widely used coordinate system, and the GENCODE annotation provides standardized gene models necessary for transcript and CDS derivation (Frankish et al., 2023). This document details a reproducible workflow to acquire, index, and operationalize hg38 for downstream analysis. Emphasis is placed on PRNP and NEFL, genes of significance in prion disease and neurofilament biology, respectively.

## Objectives

1. Build a clean local reference environment for human genomics.
2. Convert public datasets into indexed, analyst-ready artifacts.
3. Derive targeted gene assets for neuroscience/rare-disease research.
4. Deliver a reproducible, OJT-ready protocol with troubleshooting.

## Methods

### Environment and Toolchain

The workflow was executed on macOS (zsh) with the Homebrew package manager. Required tools include:

- samtools 1.22.1
- seqkit 2.10.0
- bedtools 2.31.1

- gffread 0.12.7
- pigz 2.8
- wget 1.25.0

## Reference Datasets

Genome: UCSC hg38 primary assembly (FASTA).

Annotations: GENCODE v46 basic gene annotation (GTF).

## Procedures

### System Preparation & Data Acquisition

```
brew install samtools seqkit bedtools gffread pigz wget
mkdir -p ~/genomics/hg && cd ~/genomics/hg
wget https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz -O hg38.fa.gz
pigz -d hg38.fa.gz && mv hg38.fa reference.fa
samtools faidx reference.fa
samtools dict reference.fa -o reference.dict
curl -L -o gencode.v46.basic.annotation.gtf.gz \
  https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_46/gencode.v46.basic.annotation.gtf.gz
pigz -d gencode.v46.basic.annotation.gtf.gz
mv gencode.v46.basic.annotation.gtf annotation.gtf
```

### Transcriptome & CDS FASTA

```
gffread annotation.gtf -g reference.fa -w transcripts.fa
gffread annotation.gtf -g reference.fa -x cds.fa
```

### Gene-Focused Extraction (PRNP, NEFL)

```
awk '$0 ~ /gene_name "PRNP"/' annotation.gtf > prnp.gtf
awk '$0 ~ /gene_name "NEFL"/' annotation.gtf > nefl.gtf
gffread prnp.gtf -g reference.fa -w prnp.transcripts.fa
gffread prnp.gtf -g reference.fa -x prnp.cds.fa
gffread nefl.gtf -g reference.fa -w nefl.transcripts.fa
gffread nefl.gtf -g reference.fa -x nefl.cds.fa
```

### Promoter Windows (±1 kb)

```
cut -f1,2 reference.fa.fai > genome.sizes
awk '$3=="transcript"{split($9,a,";");...}' prnp.gtf > prnp.transcripts.bed
awk '$3=="transcript"{split($9,a,";");...}' nefl.gtf > nefl.transcripts.bed
bedtools slop -i prnp.tss.bed -g genome.sizes -b 1000 > prnp.promoter_2kb.bed
bedtools slop -i nefl.tss.bed -g genome.sizes -b 1000 > nefl.promoter_2kb.bed
bedtools getfasta -fi reference.fa -bed prnp.promoter_2kb.bed -fo prnp.promoter_2kb.fa -s -name
bedtools getfasta -fi reference.fa -bed nefl.promoter_2kb.bed -fo nefl.promoter_2kb.fa -s -name
```

## Results

- PRNP: 4 transcripts, 4 CDS (762 bp; ~254 aa).

- NEFL: 1 transcript, 1 CDS (1,632 bp; ~544 aa).

- Promoter windows: PRNP (n = 4), NEFL (n = 1).

- Motif reconnaissance (±1 kb): PRNP showed 8 TATA-like motifs; NEFL showed none in this window.

## Discussion

The workflow demonstrated reliable construction of reference-aligned assets for hg38 using publicly available datasets. PRNP and NEFL extractions matched expected gene models from GENCODE v46, validating pipeline integrity. Promoter scans illustrate how lightweight motif reconnaissance can be integrated into rapid-deployment genomic analysis. Extensions include CpG island mapping, transcription factor binding site prediction (JASPAR), and variant annotation overlays.

## Risks and Troubleshooting

- Contig mismatches: Ensure chromosome naming consistency (e.g., “chr1” vs “1”).

- zsh quirks: Inline # or parentheses can cause parse errors.

- Compute constraints: Laptop environments require sufficient SSD space (>20 GB).

## Governance and Licensing

All datasets (UCSC hg38, GENCODE v46) are open for research and educational use. Attribution is required per original providers.

## References

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., … & Flicek, P. (2023). GENCODE 2023: reference annotation for the human and mouse genomes. Nucleic Acids Research, 51(D1), D942–D949. https://doi.org/10.1093/nar/gkac1072

UCSC Genome Browser. (2023). Genome downloads. https://hgdownload.soe.ucsc.edu/

## Author’s Note

*“Nobody gave me a chance, so I gave myself a chance — and AI is assisting me, 100% application and implementation, until I fully understand the process. This document is both methodology and statement of self-reliance.”*