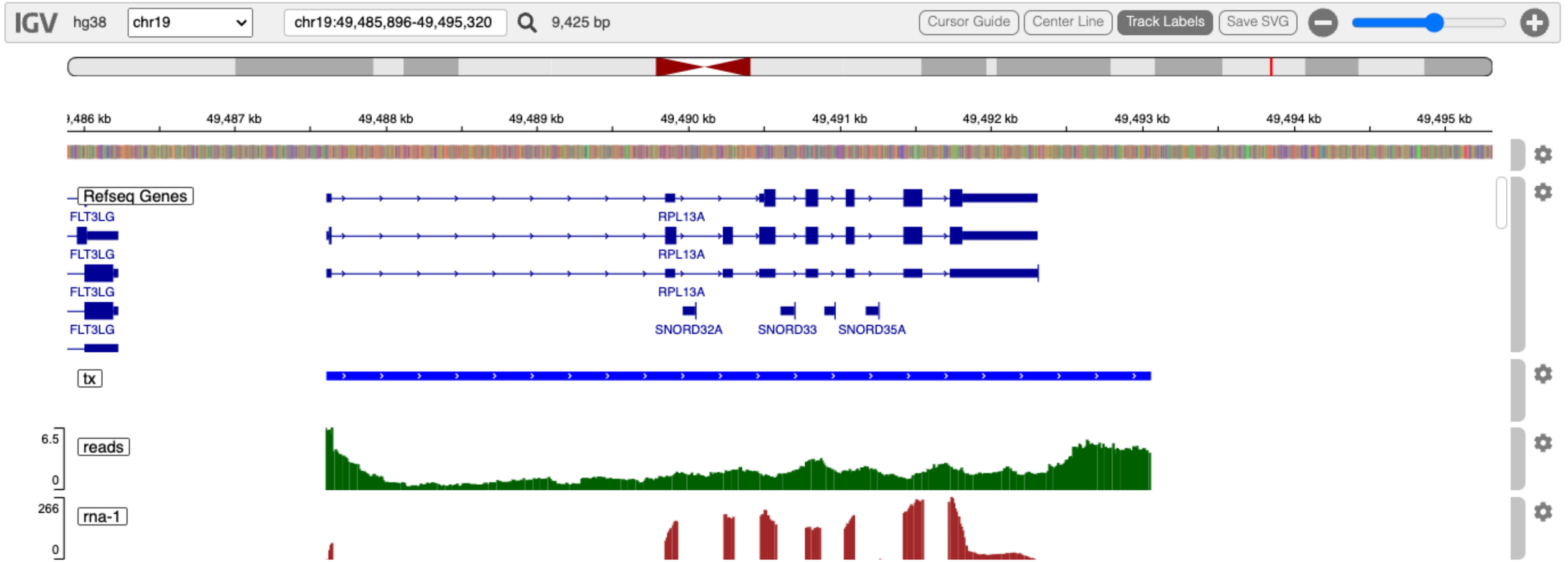
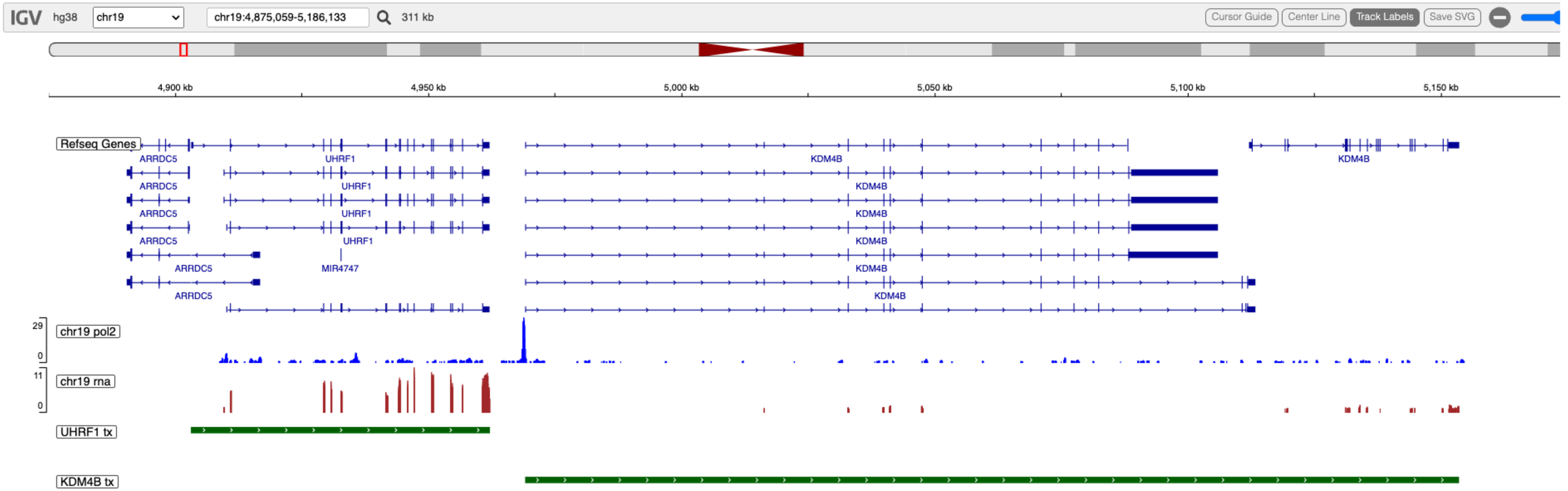
Pol2 Cut&Tag

JURKAT highly-expressed RPL13A case study: which pol2/ser2 reads should we count?



- pol2/ser2 reads in green: we likely want to ignore the pileup near the promoter, and after the 3' UTR
- proposed: count pol2/ser2 binding starting at tx transcript, go to end of
  - last expressed exon?
  - beginning of 3' UTR?
  - some location suggested by the nascent RNA study Marjorie alerted us too?

The Heterogeneity of pol2/rna relations as seen in 2 additional JURKAT genes: UHRF1 and KDM4B



- UHRF1 is relatively highly expressed (RNA-seq red, transcript in green, pol2/ser2 in blue, no pileup evident at promoter
- KDM4B is less expressed, with moderate pol2/ser2 and a pileup at promoter

ChRO-seq

Published in final edited form as:
Nat Genet. 2018 November ; 50(11): 1553–1564. doi:10.1038/s41588-018-0244-3.

Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme

Tinyi Chu[1,2], **Edward J. Rice**[1,3], **Gregory T. Booth**[4], **H. Hans Salamanca**[5], **Zhong Wang**[1], **Leighton J. Core**[6], **Sharon L Longo**[7], **Robert J. Corona**[8], **Lawrence S. Chin**[7], **John T. Lis**[4], **Hojoong Kwak**[4,*], and **Charles G. Danko**[1,3,*]

email to Jeff & Marjorie (24 apr 2023)

ChRO-seq's benefit over PRO-seq appears to be its ability to rescue information from degraded nascent RNA - for instance, from a 30-year old GBM tissue sample. The two methods produce essentially identical results with JURKAT cells.

The paper focuses first on MIR181A1, before moving to GBM:

> The microRNA MIR181 locus illustrates the advantages of ChRO-seq compared with other molecular assays (Fig. 1d). Notably, both ChRO-seq and PRO-seq discovered the primary transcript encoding MIR181 as well as dozens of eRNAs that were not discovered using RNA-seq.

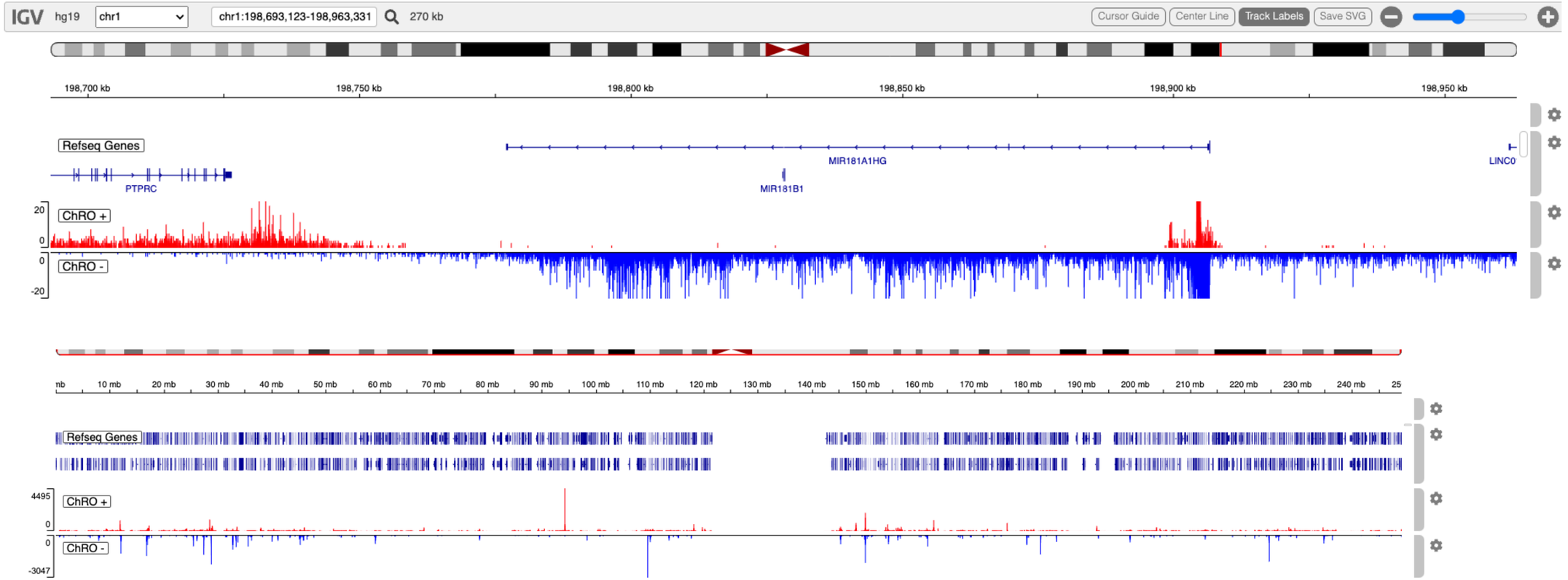
I am puzzled by their use of read threshold. The use +/- 20 (+strand and -strand respectively) - whereas the full range of read magnitude is -3047 to 4495.

They must have reason to value these relatively low values. The full chromosome, with autoscaling, is shown below. And an intermediate scaling threshold below that.

I think we can learn a lot about active enhancers from either PRO-seq or ChRO-seq, especially if we observe changes in them over the developmental trajectory.

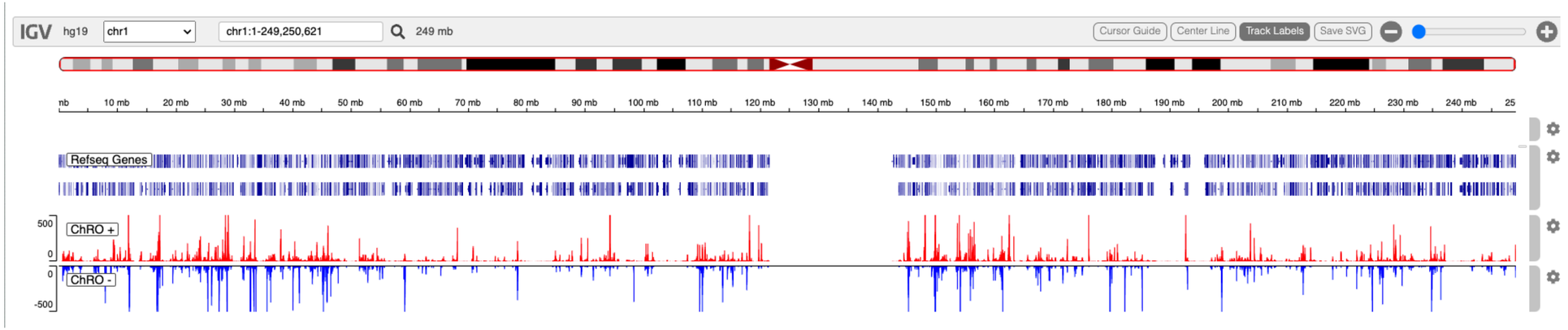
Very different activity implied by different y-axis scales. chr1. hg19.

-20:20



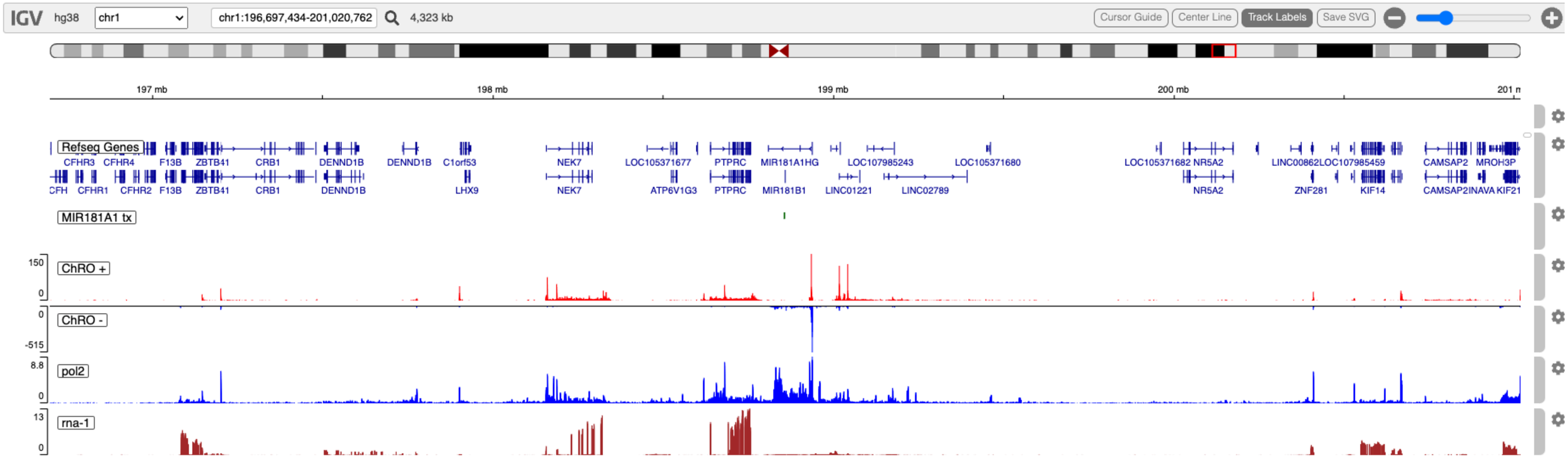
-3047:4495
(autoscale)

-500:500



Jeff points out that PRO- and ChRO-seq are too expensive, and asks: can we see any relationship between ChRO-seq and our pol2ser2 CUT&TAG? Enough to use C&T to identify active enhancers?
I ask: intergenic C&T may do so, we can check for correspondence or disagreement between C&T and ChRO-seq.

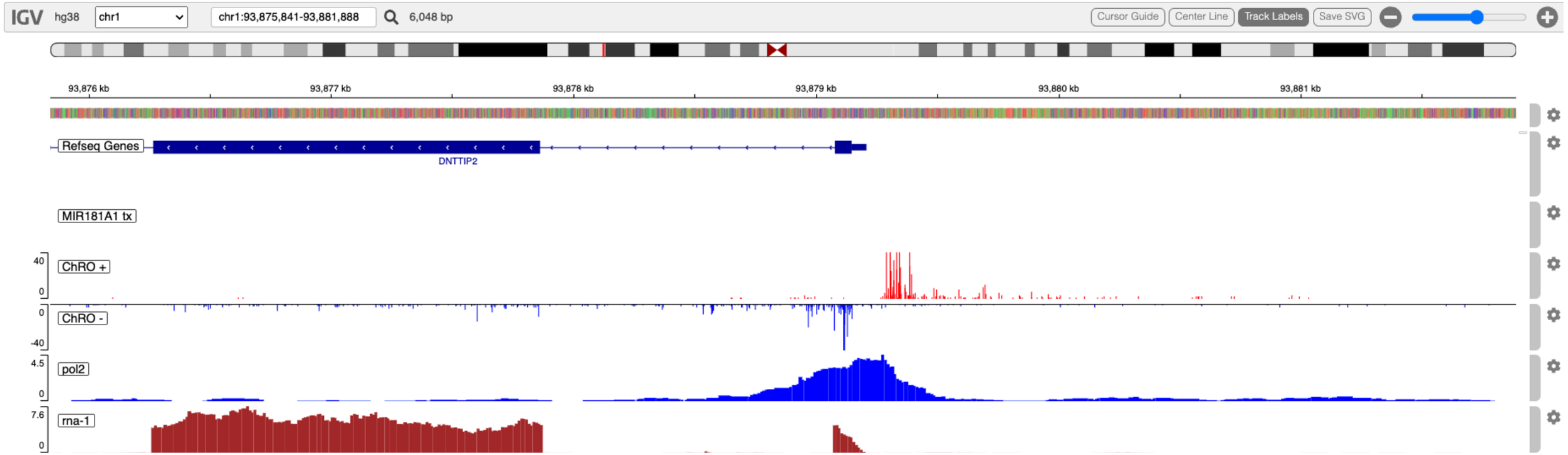
This slide shows fair alignment between ChRO and pol2ser2.



compare intergenic pol2ser2 and ChRO-seq

perhaps offering insight into the question, can C&T pol2ser2 identify active enhancers?

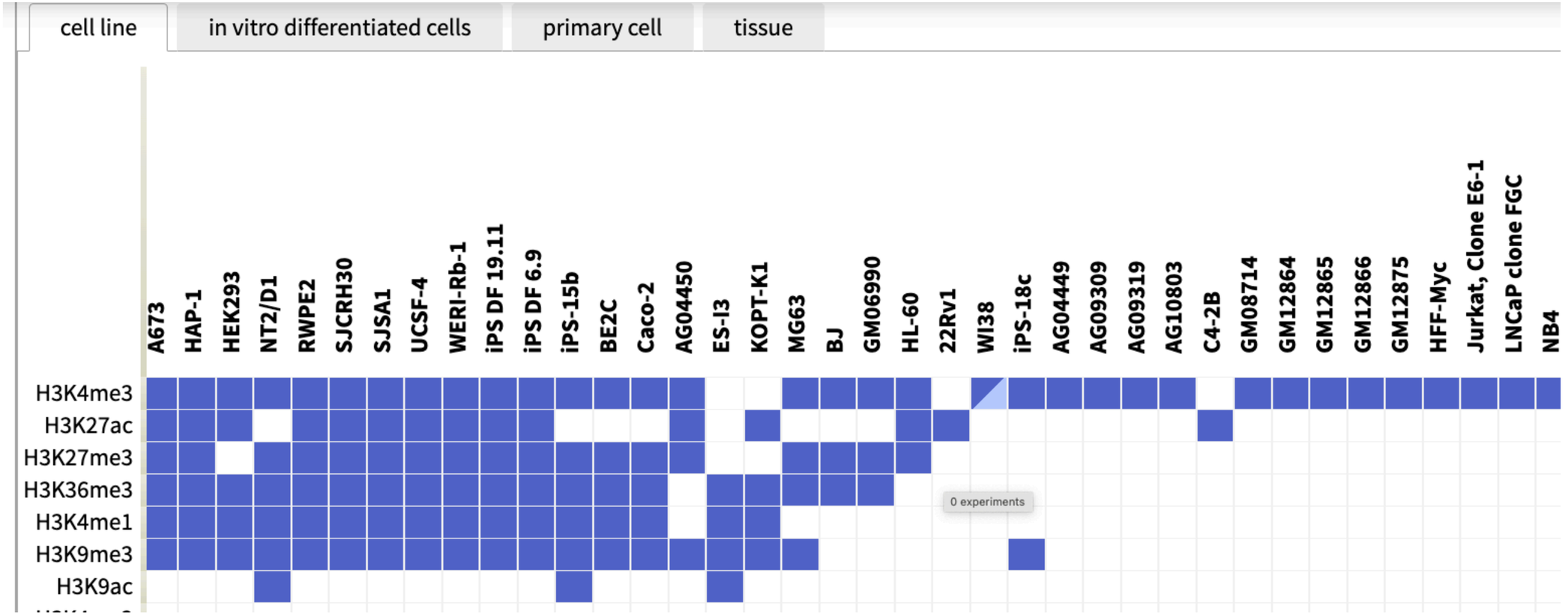
Maybe the low level pol2ser2 reads (in blue) upstream of the promoter and first exon of DNTTP2, somewhat matched to the +strand ChRO-seq reads (in bright red, suggest enhancer activity.



Looking for Jurkat H3K27Ac

ENCODE ChIP-seq matrix

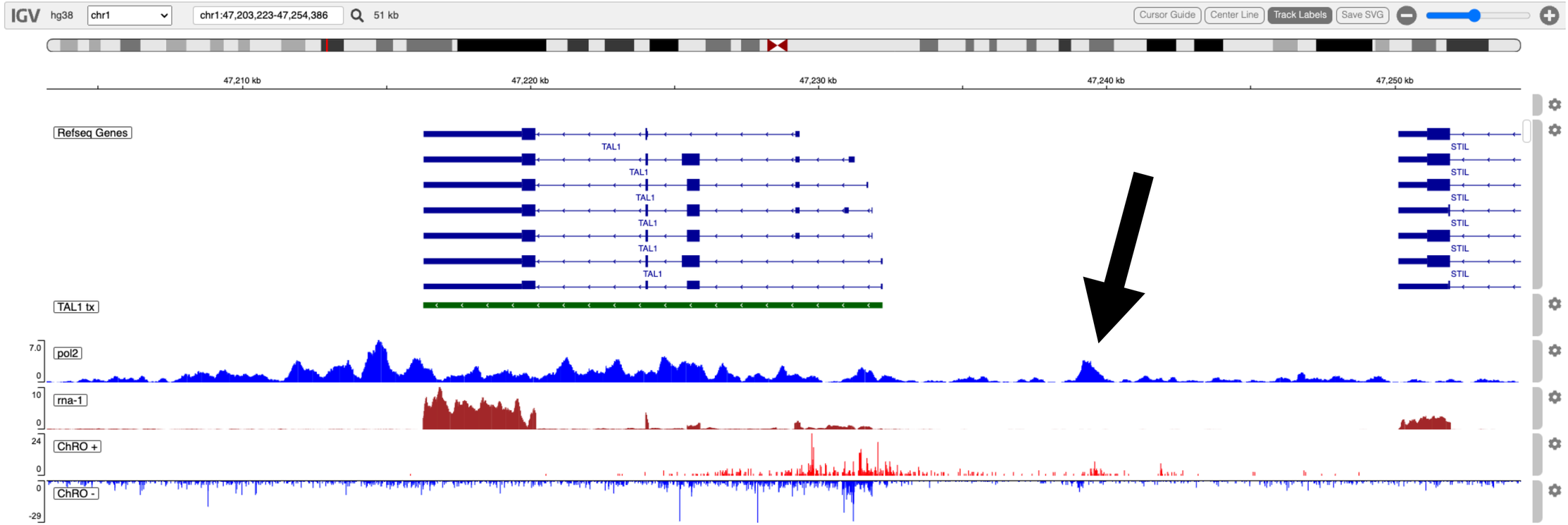
https://www.encodeproject.org/chip-seq-matrix/?type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Homo%20sapiens&assay_title=Histone%20ChIP-seq&status=released



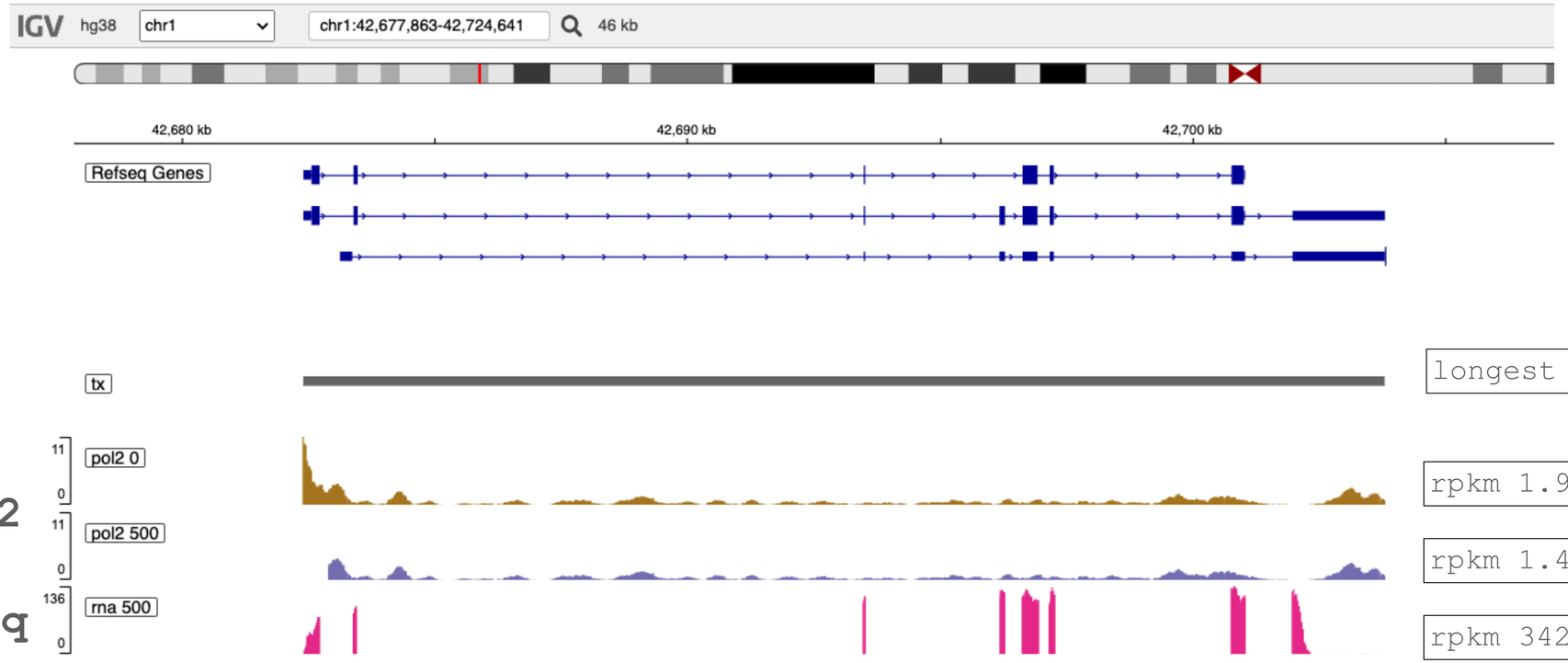
TAL1 super-enhancer reported in Mansour 2014

An Oncogenic Super-Enhancer Formed Through Somatic Mutation of a Noncoding Intergenic Element

Strong pol2ser2, negligible mRNA



YBX1: an arbitrarily selected highly-expressed gene on chr1



mRNA but no Pol2? TFs degrade quickly

In our series of experiments, we determined thousands of decay rates for transcripts in human cells. Our estimated **median mRNA half-life in human cells is 10 h**, a number that scales linearly relative to division time when compared with bacteria and yeast (Bernstein et al. 2002; Wang et al. 2002). And, we found that the number of t**ranscripts with short half-lives (i.e., <2 h) is ~5%** in both a cancer cell line (HepG2) and a primary cell line (Bud8).

[We found] statistically significant organizational principles in the variation of decay rates among functional classes. In particular, **transcription factor mRNAs have increased average decay rates compared with other transcripts and are enriched in "fast-decaying" mRNAs with half-lives <2 h. In contrast, we find that mRNAs for biosynthetic proteins have decreased average decay rates and are deficient in fast-decaying mRNAs.** Our analysis of data from a previously published study of *Saccharomyces cerevisiae* mRNA decay shows the same functional organization of decay rates, implying that it is a general organizational scheme for eukaryotes. Additionally, we investigated the dependence of decay rates on sequence composition, that is, the presence or absence of short mRNA motifs in various regions of the mRNA transcript. Our analysis recovers the positive correlation of mRNA decay with known AU-rich mRNA motifs, but we also uncover further short mRNA motifs that show statistically significant correlation with decay. However, we also note that none of these motifs are strong predictors of mRNA decay rate, indicating that the regulation of mRNA decay is more complex and may involve the cooperative binding of several RNA-binding proteins at different sites.

Genome Res. 2003 Aug; 13(8): 1863–1872. PMCID: PMC403777
doi: 10.1101/gr.1272403 PMID: 12902380

Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes

Edward Yang,[1,6] Erik van Nimwegen,[4,6] Mihaela Zavolan,[2] Nikolaus Rajewsky,[5] Mark Schroeder,[2] Marcelo Magnasco,[3] and James E. Darnell, Jr[1,7]

Hypothesis: Based on half-life data, we predict mRNA from TF genes (pearson) correlates to pol2.500 better than mRNA from non-TF genes

TFs are 1430 genes annotated to "DNA-binding transcription factor activity"
Only 109 of these are in our JURKAT data.

Somewhat promising result: better Pol2/rna correlation for TFs

```
cor.method <- "pearson"
with(subset(tbl, gene %in% tfs.all), cor(pol2.500, rna))
with(subset(tbl, ! gene %in% tfs.all), cor(pol2.500, rna))
```

```
0.270      TFs
0.132  non-TFS
```

same test as above, but use pol2.0: **0.262**
0.174

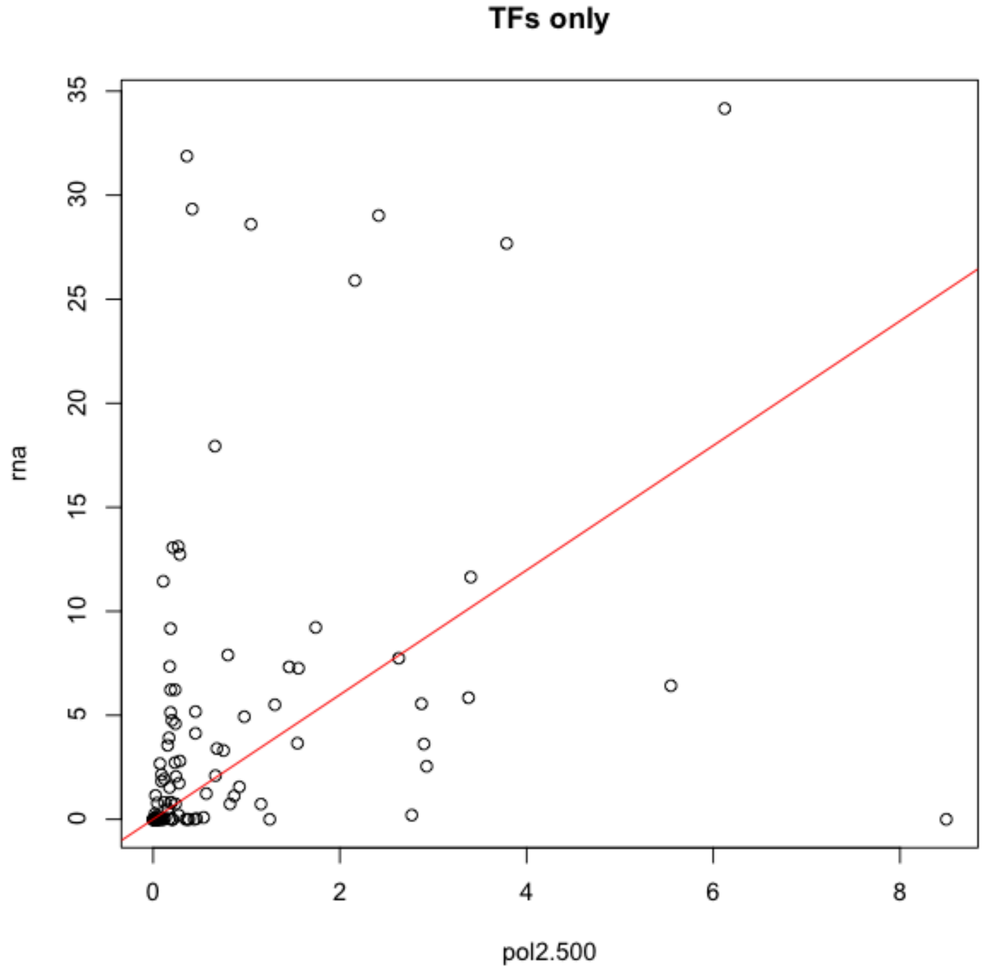
```
with(subset(tbl, strand=="+" & gene %in% tfs.all), cor(chroPlus.500,  rna))
with(subset(tbl, strand=="+" & ! gene %in% tfs.all),cor(chroPlus.500, rna))
```

```
0.282   TFs
0.257  non-TFS
```

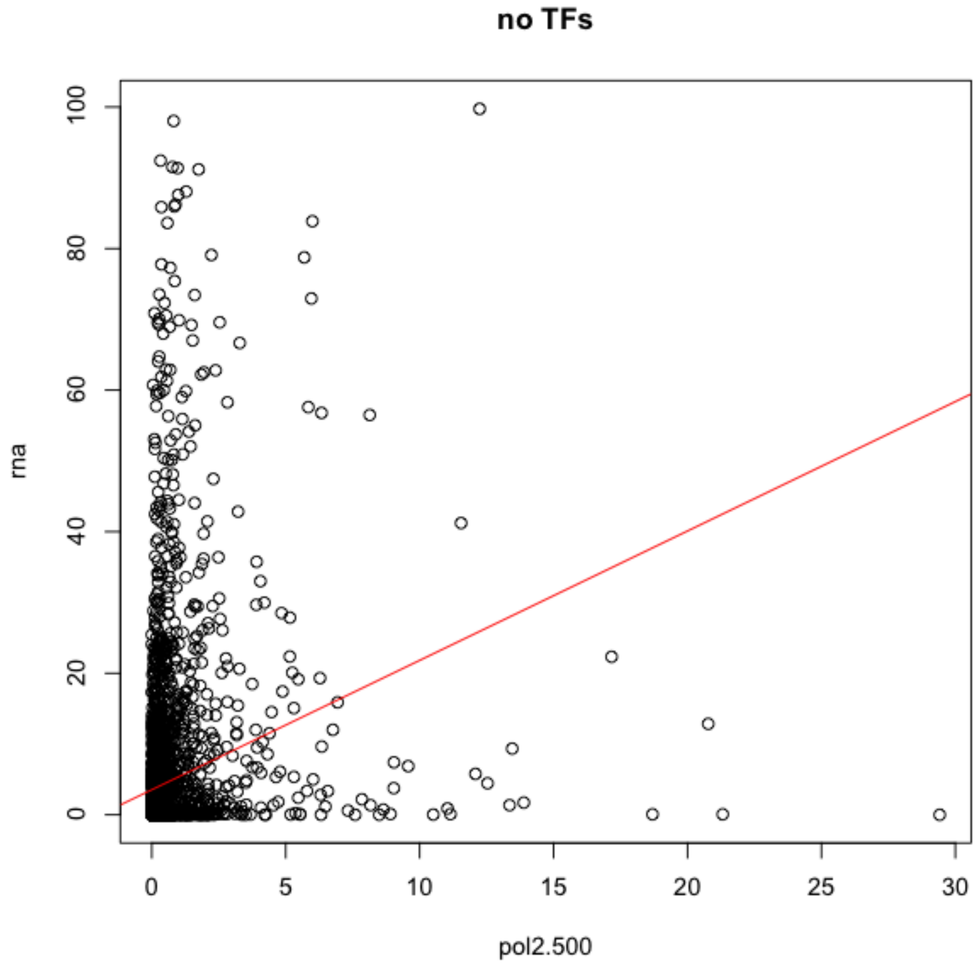
```
with(subset(tbl, strand=="-" & gene %in% tfs.all), cor(chroMinus.500, rna))
with(subset(tbl, strand=="-" & ! gene %in% tfs.all),cor(chroMinus.500, rna))
```

```
-0.006  TFs
 0.121  non-TFS
```

Omitting very highly-expressed genes, how does pol2.500 predict mRNA?



R^2: 0.293
p-value: 1.283e-09



R^2: 0.045
p-value: 2.2-16

```
tbl.filtered <- subset(tbl, gene %in% tfs.all & rna < 100)
```

Tasks from 05/09/23 zoom meeting

Hypothesis: significant pol2.500 values may change across timepoints, allowing us to say

"***our pol2 data is a better indication of active transcription than rna-seq***"

Using pol2 ChRO-seq and CUT&Tag assays, and rna-seq, all from Jurkat cells, we try to gain some preliminary insight in support of this hypothesis. Some questions arise:

1. Our ChRO-seq and CUT&Tag data measure Pol2Ser2 activity in ostensibly identical cells. Yet they come from different labs (Danko 2018, Brand Lab 2023). How do we interpret discrepancies?
2. Perhaps these two methods correlate best when RPKM values are high in both assays? And/or filtered by transcript lengths? Do we eliminate all genes that do not correlate?
3. Transcripts with a long half-life should correlate poorly with measures of active transcription. With half-life data, this "should correlate poorly" claim could be tested, and the utility of Pol2Ser2 CUT&Tag established. We don't have that data. How strong can our conclusions be? (See slide 12 for some indirect evidence of transcript half-life.)

1. identify poised transcripts by comparing first 500 bp against remainder, in both assays, using multiple strategies for filtering to identify strong candidates (see 3 and 4 below).
2. exclude anti-sense genes
3. does minimum transcript length affect pol2/ChRO-seq/rna correlation? we currently filter out transcripts < 1000 bp. try some other values.
4. Same as 3, using RPKM filtering. Consider RPKM ChRO, RPKM C&T, then RPKM > filter in both assays. Then:
5. find genes where pol2 c&t disagrees with ChRO-seq, make genomic figure with tracks
6. create new column: sum of chro-seq plus and minus, 0 and 500, correlate with pol2 0 and 500
7. find genes where x.500 is greater than x.0: make genomic figure with tracks.
8. after slides summarizing tasks 1-7 are reviewed, do whole genome.

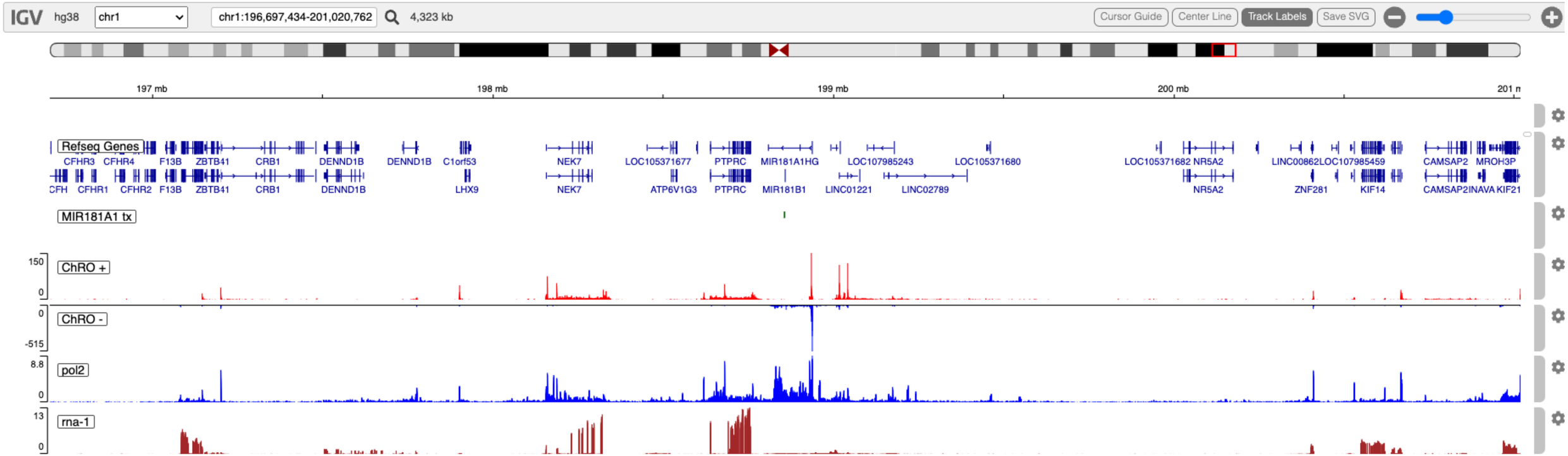
Jeff replies - email (1 jun 2023):

The tasks look good to me. What I would be most interested to see is:
1) the correlation between pol II C&T and Chro seq.
2) Optimize the correlation between pol II C&T and Chro-seq . I think tasks 3 and 4 are in line with this point.

Jurkat Pol2 CUT&Tag compared to ChRO-seq
chr1 only

	pol2.0 vs chroBoth.0	pol2.500 vs chroBoth.500	pol2.500 vs chroPlus.500	pol2.500 vs chroMinus.500	gene count
all chr1	0.24	0.32	0.32	0.20	3808
rna pkrm > 0.5	0.28	0.43	0.43	0.21	1194
rna > 0.5, tx > 1000	0.28	0.44	0.44	0.21	1179
rna > 0.5 tx < 1000	0.89	0.24	0.24	0.30	15
rna > 5.0 tx > 1000	0.25	0.39	0.39	0.33	757
rna > 5.0 tx > 10000	0.32	0.34	0.34	0.25	620
rna > 0.5, tx >1000 AS genes dropped	0.27	0.40	0.40	0.22	1137

Jeff points out that this 4.3M region of chr1 appears to have pretty good correlation between pol2 and ChRO+ combined with ChRO-



In this 4.32M region	
non-zero pol2 aligned reads:	39,230
non-zero ChRO aligned reads:	38,600 (+-)
shared:	32,719
pol2 shared/total:	83%
ChRO shared/total:	85%

All chr1	
non-zero pol2 aligned reads:	1,653,349
non-zero ChRO aligned reads:	1,105,548
shared:	918,926
pol2 shared/total:	56%
ChRO shared/total:	83%

Full Genome	
non-zero pol2 aligned reads:	16,741,128
non-zero ChRO aligned reads:	11,130,074
shared:	9,098,029
pol2 shared/total:	54%
ChRO shared/total:	82%

correlations with
rpkm re-expressed as quintiles

	rna vs chroBoth.0	rna vs. chroBoth.500	rna vs. pol2.0	rna vs.pol2.500	chrBoth.0 vs pol2.0	chroBoth.500 vs pol2.500
pearson	0.69	0.68	0.59	0.57	0.76	0.73
spearman	0.73	0.73	0.62	0.59	0.76	0.73
tx >1000 pearson	0.70	0.69	0.59	0.60	0.76	0.73
tx > 1000 spearman	0.74	0.73	0.62	0.57	0.76	0.73