


**From:** Fan, Jean jeanfan@fas.harvard.edu   
**Subject:** Re: Question regarding the GEO data (THS-seq) from your recent Nat. Biotech paper  
**Date:** February 5, 2018 at 9:55 AM  
**To:** Yungil Kim ipw012@gmail.com, "Brandon C Sos"  
**Cc:** "Lake, Blue", Jerold Chun jchun@sbpdiscovery.org, Kharchenko, Peter Peter\_Kharchenko@hms.harvard.edu, "Zhang, Kun-forward"

JF

Hi Yungil,

Someone else asked the same question a few days ago, so please find below the reply from Dr. Brandon Sos (cc-ed), who developed the scTHS-seq method and generated the scTHS-seq data.

----

Hi Adriana,

I've attached some examples of the input files so hopefully that should help, and the perl script used to make the split config files.

So deindexer can't demultiplex more than ~1000 indexes in a single instance for single end reads (we run 960 indexes at a time). So we split the CFG file into config files with 960 lines each using the perl script. For paired end it can't demultiplex more than ~500 indexes, and we use ~450.

We run all the deindexer instances in the background, and it takes into account the number of config files. We used to have different numbers of total indexes used, thus different numbers of split config files but now we use one index set. The script retrieves the index file names, generates a script that runs 2 config files per job, then runs it in the background.

Shell script used (need to put in your own path):

```
#=====
echo "copied over config files, running deindexer"

#Store config file names into array. array lengths determines how many processes are
run
filearray=(scTHS.cfg_*)

job_num=1
for (( i = 0 ; i < ${#filearray[@]} ; i=$i+2 ));
do
    echo "for f in ${filearray[i]} ${filearray[i+1]}" > job_$job_num.sh
    echo "do" >> job_$job_num.sh
    echo "mkdir out_\$f" >> job_$job_num.sh
    echo "/PATH_TO_DEINDEXER/deindexer -f \"rbbb\" -c \$f -b i7_scTHS -b i5_scTHS
-b r5_scTHS \">> job_$job_num.sh
    echo " Undetermined_S0_L001_R1_001.fastq Index7.i1.fastq Index5.i1.fastq
Index5.i2.fastq -o out_\$f -m $num_mismatches" >> job_$job_num.sh
    echo "done">> job_$job_num.sh
    sh job_$job_num.sh 2> job_$job_num.sh.out &
    job_num=$((job_num+1))
done
```

wait

#=====

Any questions let me know, hope this helps.

Best,  
Brandon

---

Dear Jean,

I've been following your work on single-cell analysis in the brain and came across your paper published in December 2017 (Nat Biotech 36, 70-80 (2018)) and I've been trying to replicate the data analysis performed for scTHS-seq data processing.

However, I am having trouble demultiplexing the data, for which I hope you or someone from your team might be able to help me.

My understanding is that from the string:  
AAGAGGCA\_AAGGAGTA\_AACACG\_R10P5P6, the first sequence is the i7 barcode, i5 comes second, followed by the r5 transposon barcode. When I create a cfg file, I can't make the deindexer master to work, so I am guessing I am feeding the data incorrectly.

I hope you could provide me with the in-house perl scripts that you used for this analysis and or maybe an example of how the cfg file looks like.

Thank you for your help with this matter.

Best regards,

**Adriana Seden-Cortes**  
Research Associate I



[www.alleninstitute.org](http://www.alleninstitute.org)

On Feb 2, 2018, at 9:34 PM, Yungil Kim <[ipw012@gmail.com](mailto:ipw012@gmail.com)> wrote:

Dear Dr. Lake,

I have questions regarding your THS-seq data in your recent Nat. Biotech. paper. When I downloaded your THS-seq data from GEO (SRP104139), their fastq files were merged into a list of merged ones without barcode information.

Your "online methods" section explained them as "Barcode combinations associated with each read were appended to each read header with in-house Perl scripts, and all FASTQ files were combined and mapped to an hg38 no-alternative loci plus decoy reference genome (GCA\_000001405.15\_GRCh38\_no\_alt\_plus\_hs38d1\_analysis\_set) and mm10 no-alternative loci reference genome (GCA\_000001635.5\_GRCm38.p3\_no\_alt\_analysis\_set) using BWA 0.7.12-r1039. Mapped SAM outputs were re-demultiplexed by barcode and converted to BAM files, and clonal reads were removed with SAMtools 1.3.1 while

here is demultiplexed by barcodes and converted to BAM files, and chimeric reads were removed with chimera-filt, while read statistics were gathered for each barcode combination. ".

Could you please explain how I can demultiplex them? Thank you so much.

Sincerely,  
Yungil



scTHS.CFG



scTHS.CFG\_AA



scTHS.CFG\_AB



scTHS.CFG\_AC



makeDeIndexer  
Config.perl