Derived allele frequency in Human and Macaque brain promoters

Yiru GONG

2020/8/16

Overview

Aim: To analyze the evolution tendency of promoters in different brain regions in human and macaque.
Method: Compare and analyze the SNP derived allele frequency of promoters between conserved and divergent category, calculate the odds ratio.

Original Data:

1) Biased promoter expression in 15 brain regions of Human and Macaque

```
$ cd /exports/cmvm/datastore/sbms/groups/young-lab/macaque/data/
$ ls hg19.*.outcomes.bed.gz rheMac3.*.outcomes.bed.gz
```

2) Overall human promoter expression aligned with macaque promoter

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/macaque/data/hg19_expression_outcomes.bed.gz
```

3) Human SNP alleles data

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/human/hg19.decode.snpResolved.bed.gz
```

Codes:

```
$ cd /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/
```

Files:

- *brain_region.R* —— Promoter alignment annotation & whole brain file generation
- *count.sh* & *window_count.sh* —— Codes to intersecting SNP with promoter file or window version respectively
- *subset.sh* —— dividing count files into matched and divergent subsets
- *window.sh* —— Codes to apply *bedtools window* directly
- *odds_ratio.R* —— Odds ratio calculation

Output data: all files stored in the direction:

```
$ cd /exports/cmvm/datastore/sbms/groups/young-lab/yiru/data/
```

File types:

- *"hg19."* —— Human promoter file
- *"rheMac3."* —— Macaque promoter file
- *".expression.txt"* —— Annotated brain region biased promoters
- *".expression.window.txt"* —— Expanded promoter location based on "*.expression.txt"
- *".counts.txt"* —— Intersect of SNP alleles and promoter expression
- *".counts.window.txt"* —— Intersect of SNP alleles and expanded promoter expression
- *".matched.counts.window.txt"* & *".divergent.counts.window.txt"* —— Intersect files of two subsets of promoters

- *"hg19.OR...Rdata"* —— Odds ratio results of all brain regions

I. Human-Macaque promoter alignment annotation

Code:

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/brain_region.R
```

Subsets:

1. **"Brain_expression" column:** To observe the biased expression of human and macaque promoter among different regions, the promoters in every brain region are annotated with two categories under the "Brain_expression" column:

- ***"region_specific":*** The promoter is biased in only one brain region in the species (human or macaque)
- ***"non_specific":*** The promoter is biased in multiple brain regions in the species (human or macaque)

2. **"Macaque" Column:** To discover the conservation of human promoters in different brain regions through evolution, human promoters in every brain region are annotated with three categories under the "Macaque" column.

- ***"matched":*** The human promoter has ortholog macaque promoters in the same brain region
- ***"divergent":*** The human promoter has ortholog macaque promoters, but not expressed in the same brain region
- ***"unconserved":*** The human promoter without ortholog macaque promoters in the brain

Extra Information:

1. **"Macaque_Id" Column:** For all "matched" and "divergent" promoters, the corresponding promoters expressed in the brain of the other species (a direct copy from *"hg19_expression_outcomes.bed.gz" Macaque_Id column*).

2. **"matched_Id" Column:** Only for "matched" promoters, the corresponding macaque promoters expressed in the same region.

The Macaque files are annotated with the same criteria.

Output:

Finally, we get 15 files of annotated region-biased promoters for each species.

```
$ vi *.expression.txt
### Eg. hg19.amygdala.expression.txt
```

Each file contains 10 columns:

- *chr* - promoter location of chromosome
- *start* - starting locus of the promoter
- *end* - ending locus of the promoter
- *Id* - promoter Id
- *TPM* - Transcripts PerKilobase Million
- *Strand* - Strand of the sequence
- *Brain_expression* - annotation with two category "region_specific/non_specific"
- *Macaque* - promoter alignment annotation with three category "matched/divergent/unconserved"
- *Macaque_Id* - corresponding macaque Ids expressed in the brain
- *matched_Id* - corresponding macaque Ids expressed in the same region

II. Generate overall human brain promoter expression file

Code: Included in the previous annotation code.

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/brain_region.R
```

Output File:

1. hg19.all_brain.expression.txt

To compare the promoter expression between specific brain regions and overall brain expression, we put human promoters in all brain regions together and delete the repeated promoters. Every promoter is annotated in three categories:

- ***"matched":*** The promoter is ortholog to some macaque promoters in AT LEAST ONE same region
- ***"divergent":*** The promoter is ortholog to some macaque brain promoters but NONE is expressed in the same region
- ***"unconserved":*** The human promoter without ortholog macaque promoters

2. hg19.all_brain_downsampled.expression.txt

To keep the similar statistical power, average number of promoters in each region is calculated and the same number of promoters are randomly selected from the *all_brain* file. 5 repeats are generated in the subdirectory:

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/data/subsampled/
```

III. SNP allele intersection

To compare the allele frequency change of promoters in all brain regions, we intersect the human SNP expression with the promoter expression using **bedtools intersect** command.

Code:

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/count.sh

# Example.
$ module load igmm/apps/BEDTools/2.25.0
$ bedtools intersect -a hg19.decode.snpResolved.bed.gz -b
  hg19.brain_general.expression.txt -wb -sorted > hg19.brain_general.counts.txt
```

Output:

```
# suffix "*.counts.txt"
$ ls *.counts.txt
```

Calculate Odds ratio of derived allele frequency

Code:

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/odds_ratio.R
```

We first count the overall common alleles (derived allele frequency **>5%**) and rare alleles (derived allele frequency **<1.5%**) of all SNP (*"hg19.decode.snpResolved.bed.gz"*).

```
$ zcat hg19.decode.snpResolved.bed.gz | awk '{print $9}' | awk '$1<0.015{print}' | wc -l
```

```
rare_genome    = 12205724
common_genome  = 5263405
```

For each counts file, we count the number of common and rare alleles, and perform Fisher exact test to

calculate the odds ratio, P-value and confidential interval.

	Region	*Genome*
rare		12205724
common		5263405

$$Odds.Ratio = \frac{rare.region/common.region}{rare.genome/common.genome}$$

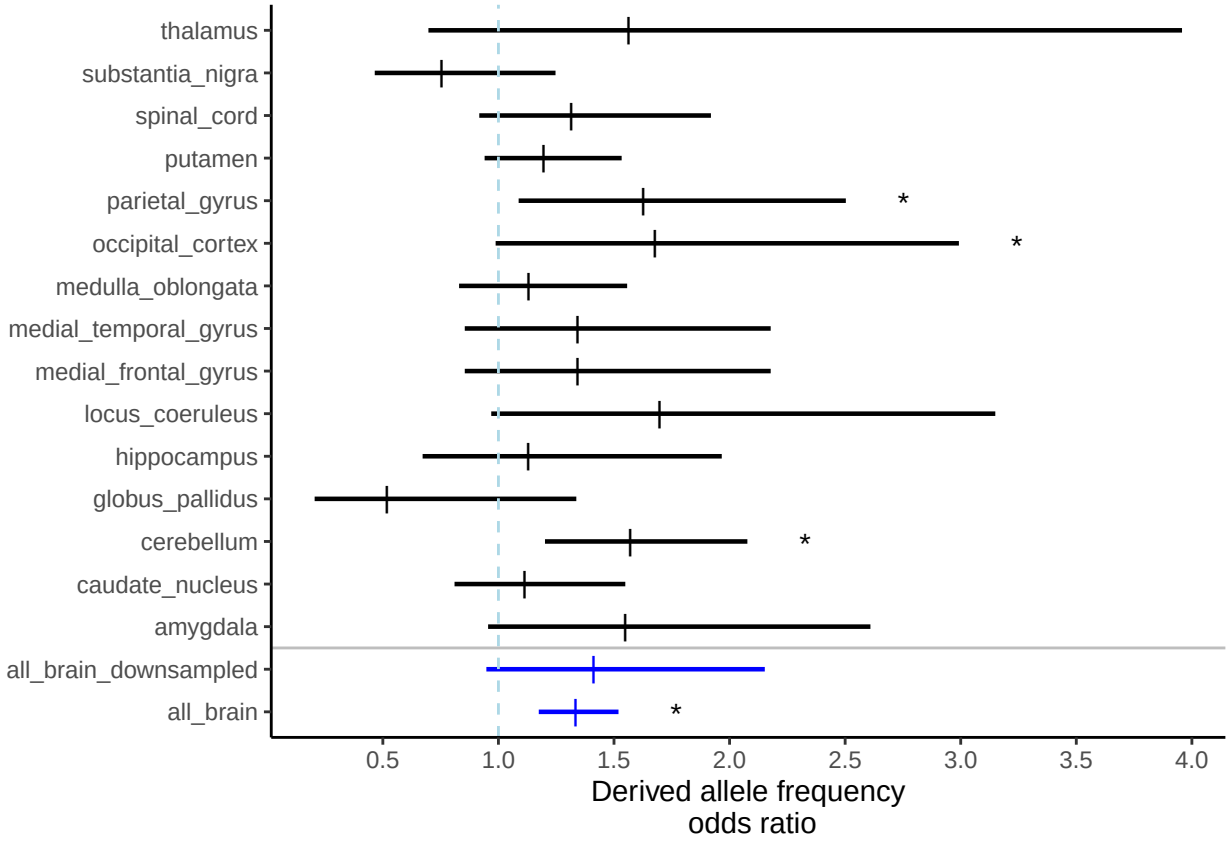
Result:

```
load('C:/Users/lenovo/Desktop/Young lab/data/hg19.OR.Rdata')
```

	brain_region	rare_region	common_region	rare_genome	common_genome
2	all_brain	981	317	12205724	5263405
1	all_brain_downsampled	108	33	12205724	5263405
3	amygdala	79	22	12205724	5263405
6	caudate_nucleus	142	55	12205724	5263405
7	cerebellum	255	70	12205724	5263405
8	globus_pallidus	12	10	12205724	5263405
9	hippocampus	55	21	12205724	5263405
10	locus_coeruleus	63	16	12205724	5263405
11	medial_frontal_gyrus	81	26	12205724	5263405
12	medial_temporal_gyrus	81	26	12205724	5263405
13	medulla_oblongata	152	58	12205724	5263405
14	occipital_cortex	70	18	12205724	5263405
15	parietal_gyrus	117	31	12205724	5263405
16	putamen	258	93	12205724	5263405
17	spinal_cord	125	41	12205724	5263405
18	substantia_nigra	49	28	12205724	5263405
19	thalamus	29	8	12205724	5263405

	brain_region	odds_ratio	lower_conf	upper_conf	p.value
2	all_brain	1.3344394	1.1745606	1.519516	0.0000055
1	all_brain_downsampled	1.4113322	0.9478733	2.152356	0.0979917
3	amygdala	1.5485083	0.9553322	2.609337	0.0819809
6	caudate_nucleus	1.1133400	0.8101071	1.548966	0.5349788
7	cerebellum	1.5709037	1.2014765	2.077010	0.0005578
8	globus_pallidus	0.5174457	0.2048951	1.337039	0.1603114
9	hippocampus	1.1293931	0.6720551	1.966166	0.7084294
10	locus_coeruleus	1.6979455	0.9690475	3.149643	0.0651156
11	medial_frontal_gyrus	1.3433835	0.8543847	2.177872	0.2068844
12	medial_temporal_gyrus	1.3433835	0.8543847	2.177872	0.2068844
13	medulla_oblongata	1.1301011	0.8299220	1.556938	0.4527560
14	occipital_cortex	1.6769833	0.9878933	2.992178	0.0485057
15	parietal_gyrus	1.6275239	1.0874917	2.502955	0.0152866
16	putamen	1.1962879	0.9402996	1.533179	0.1459453
17	spinal_cord	1.3146706	0.9172140	1.919414	0.1500541
18	substantia_nigra	0.7546250	0.4650099	1.246993	0.2632027
19	thalamus	1.5632053	0.6972317	3.957333	0.2883350

Plot:



Interpret:

As the result indicated, the odds ratio of the whole brain is significantly larger than 1, showing a tendency of more rare alleles, which represent a purifying selection in brain evolution. But the downsampled one do not reach the significant level. Some brain regions (**parietal gyrus, occipital cortex** and **cerebellum**) also showed significant odds ratio, while others didn't.

IV. Window intersection

Since the nearby region of the promoter sequence may also affect the function of the promoter, we additionally perform *bedtools window* to expand the intersecting region. In specific, 150 bp is added to the upstream of promoter location and 50 bp to the downstream.

Code: (regardless of storage problem)

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/window.sh

# For example:
$ bedtools window -l 150 -r 50 -sw -u -a hg19.amygdala.expression.txt -b
  hg19.decode.snpResolved.bed.gz > hg19.amygdala.counts.window.txt
```

Due to the storage limitation, we perform the addition manually in *R* and then run *bedtools intersect* as the above method.

```
filepath <- '/mnt/nfsstaging/cmvm/datastore/sbms/groups/young-lab/yiru/data'
filenames <- list.files(filepath,pattern='hg19.*.expression.txt$')
```

```
for (idx in 1:length(filenames)){
  data = read.table(filenames[idx],stringsAsFactors = F,header=T)
  data[which(data$Strand=='+'),2] = data[which(data$Strand=='+'),2]-150
  data[which(data$Strand=='+'),3] = data[which(data$Strand=='+'),3]+50
  data[which(data$Strand=='-'),2] = data[which(data$Strand=='-'),2]-50
  data[which(data$Strand=='-'),3] = data[which(data$Strand=='-'),3]+150

  data = data[order(data[,1],data[,2],data[,3]),]
  name = gsub('expression','expression.window',filenames[idx])
  write.table(data,name,row.names = F,quote=F, sep='\t')
}
```

Then the same operation of *bedtools intersect* is performed.

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/window_count.sh
```

Output:

- Window expression file:

```
# suffix "*.expression.window.txt"
$ ls *.expression.window.txt
```

- Windowed SNP intersected file:

```
# "*.counts.window.txt"
$ ls | grep .counts.window.txt | grep -v divergent | grep -v matched
```

Calculate Odds ratio of derived allele frequency

The same process is conducted as the above not windowed files.

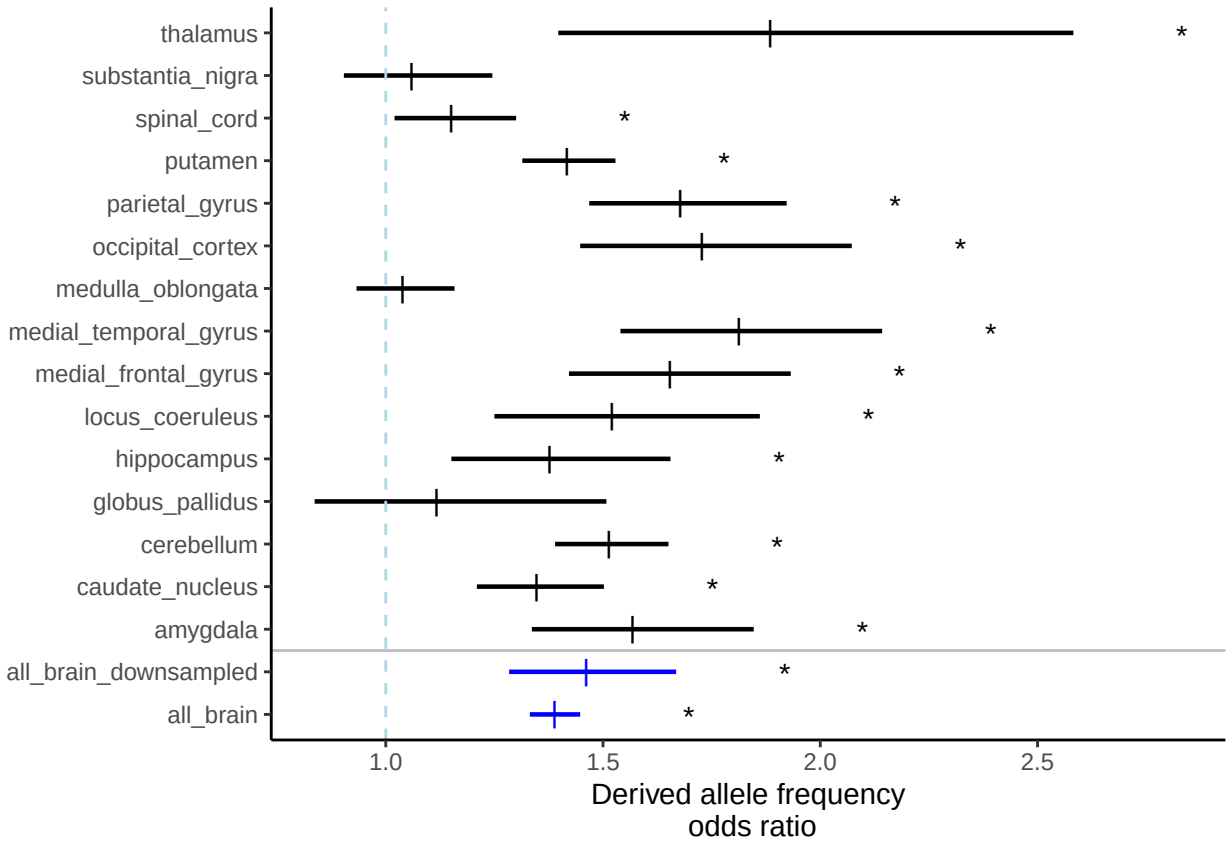
Result:

```
load('C:/Users/lenovo/Desktop/Young lab/data/hg19.OR.window.Rdata')
```

	brain_region	rare_region	common_region	rare_genome	common_genome
2	all_brain	9426	2928	12205724	5263405
1	all_brain_downsampled	1017	300	12205724	5263405
3	amygdala	709	195	12205724	5263405
6	caudate_nucleus	1390	445	12205724	5263405
7	cerebellum	2394	682	12205724	5263405
8	globus_pallidus	171	66	12205724	5263405
9	hippocampus	511	160	12205724	5263405
10	locus_coeruleus	462	131	12205724	5263405
11	medial_frontal_gyrus	821	214	12205724	5263405
12	medial_temporal_gyrus	765	182	12205724	5263405
13	medulla_oblongata	1149	477	12205724	5263405
14	occipital_cortex	625	156	12205724	5263405
15	parietal_gyrus	1070	275	12205724	5263405
16	putamen	2944	896	12205724	5263405
17	spinal_cord	990	371	12205724	5263405
18	substantia_nigra	538	219	12205724	5263405
19	thalamus	236	54	12205724	5263405

	brain_region	odds_ratio	lower_conf	upper_conf	p.value
2	all_brain	1.388281	1.3316752	1.447445	0.0000000
1	all_brain_downsampled	1.461893	1.2839533	1.668279	0.0000000
3	amygdala	1.567903	1.3362413	1.846793	0.0000000
6	caudate_nucleus	1.346923	1.2097156	1.502068	0.0000000
7	cerebellum	1.513741	1.3896970	1.650579	0.0000000
8	globus_pallidus	1.117260	0.8363421	1.507744	0.4792714
9	hippocampus	1.377166	1.1509791	1.655260	0.0003411
10	locus_coeruleus	1.520834	1.2500325	1.860879	0.0000135
11	medial_frontal_gyrus	1.654370	1.4215877	1.932049	0.0000000
12	medial_temporal_gyrus	1.812563	1.5401064	2.142427	0.0000000
13	medulla_oblongata	1.038709	0.9327284	1.158200	0.4992950
14	occipital_cortex	1.727661	1.4474523	2.072380	0.0000000
15	parietal_gyrus	1.677854	1.4682271	1.922609	0.0000000
16	putamen	1.416931	1.3142914	1.528604	0.0000000
17	spinal_cord	1.150700	1.0202531	1.300036	0.0211875
18	substantia_nigra	1.059354	0.9036499	1.245312	0.5007268
19	thalamus	1.884610	1.3970289	2.582368	0.0000119

Plot:



Interpret:

The result showed that after expanding the upstream and downstream intersection, most brain regions (except **substantia nigra, medulla oblongata** and **globus pallidus**) become significantly larger than 1, indicating a conserved purifying selection among most brain regions.

Since the window option returns more significant results, we continued with the windowed files.

V. Compare derived allele frequency between matched and divergent promoter

To further observe the odds ratio change between conserved and aligned promoters, we separate the counts file into two subsets according to the promoter category (conserved vs aligned) and count the allele frequency separately.

Code:

```
$ /exports/cmvm/datastore/sbms/groups/young-lab/yiru/code/subset.sh
```

```
# For example:
$ awk '$19=="conserved" {print}' hg19.amygdala.counts.window.txt >
  hg19.amygdala.matched.counts.window.txt
$ awk '$19=="aligned" {print}' hg19.amygdala.counts.window.txt >
  hg19.amygdala.divergent.counts.window.txt
```

Therefore, two subset counts files are generated.

Output:

```
$ ls *.divergent.counts.window.txt *.matched.counts.window.txt
```

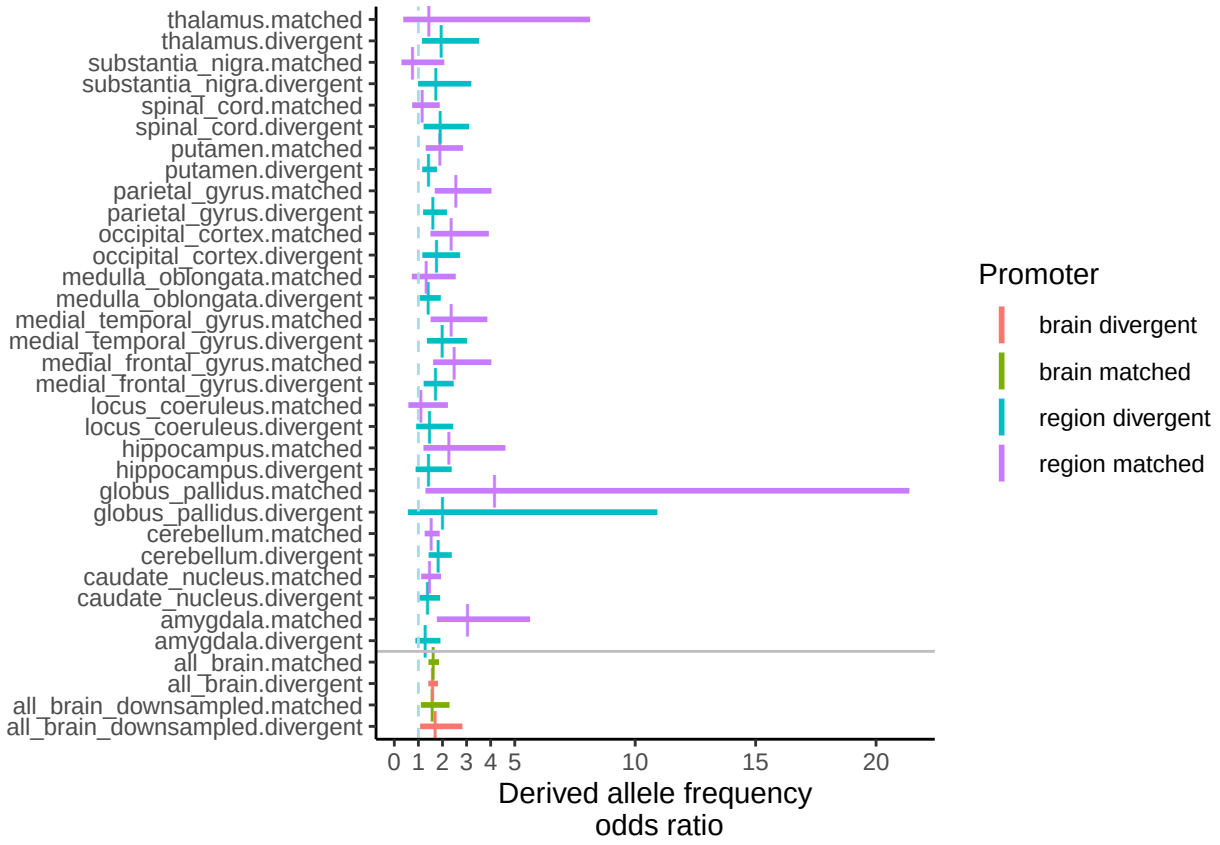
Calculate Odds ratio of derived allele frequency

applying the same method of odds ratio calculation (region/genome).

Result:

```
load('C:/Users/lenovo/Desktop/Young lab/data/hg19.OR.subsets.window.Rdata')
```

Plot:



Compare matched and divergent groups directly

To compare the matched and divergent groups more clearly, we change the background to the divergent group and calculate the odds ratio of the two groups in each brain region directly.

	Matched	*Divergent*
rare		
common		

$$Odds.Ratio = \frac{rare.matched/common.matched}{rare.divergent/common.divergent}$$

Result:

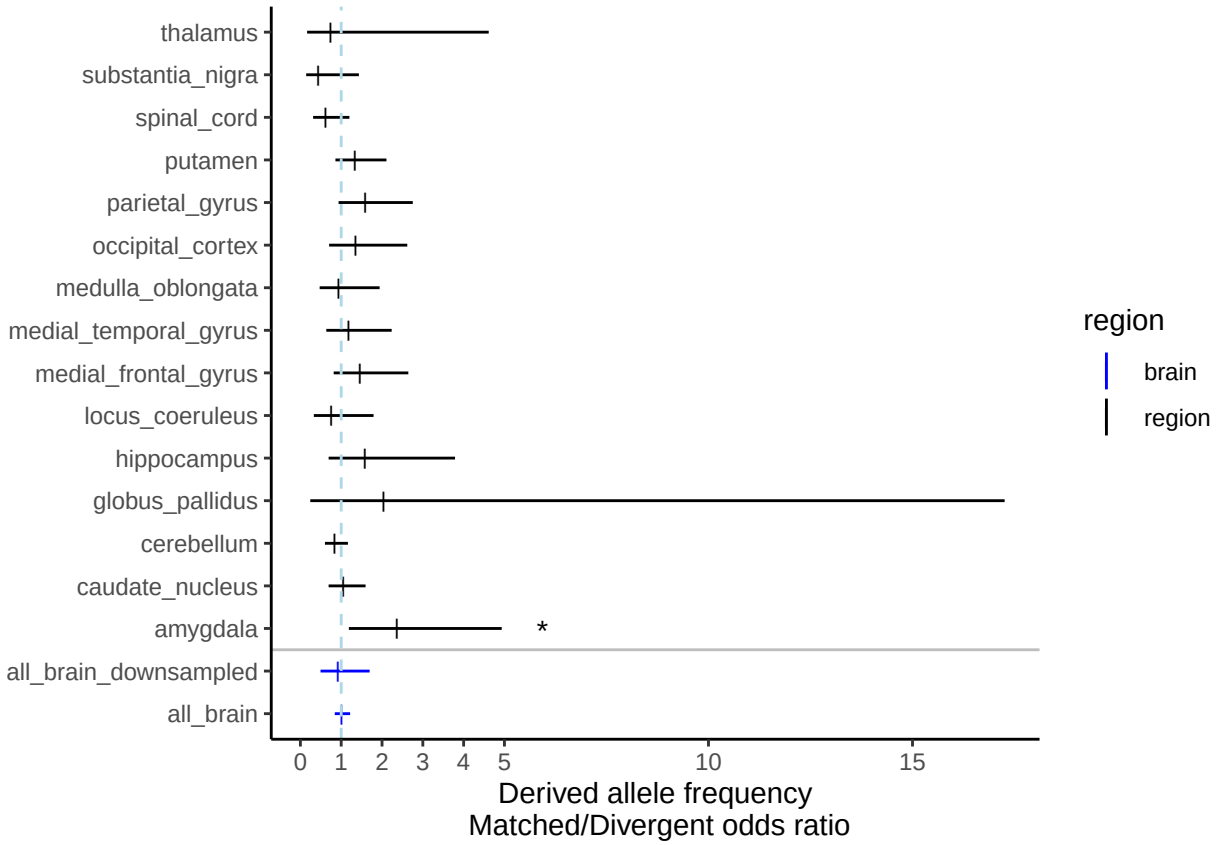
```
load('C:/Users/lenovo/Desktop/Young lab/data/hg19.OR.matched_divergent.window.Rdata')
```

	brain_region	rare_matched	common_matched	rare_divergent	common_divergent
2	all_brain	1047	279	1200	324
1	all_brain_downsampled	146	40	91	23
3	amygdala	106	15	110	37
6	caudate_nucleus	234	69	187	58
7	cerebellum	456	128	323	76
8	globus_pallidus	29	3	14	3
9	hippocampus	63	12	76	23
10	locus_coeruleus	36	14	78	23
11	medial_frontal_gyrus	139	24	167	42

	brain_region	rare_matched	common_matched	rare_divergent	common_divergent
12	medial_temporal_gyrus	126	23	148	32
13	medulla_oblongata	46	15	190	58
14	occipital_cortex	121	22	118	29
15	parietal_gyrus	154	26	216	58
16	putamen	150	34	387	117
17	spinal_cord	73	27	106	24
18	substantia_nigra	14	8	64	16
19	thalamus	10	3	77	17

	brain_region	odds_ratio	lower_conf	upper_conf	p.value	p.adjust
2	all_brain	1.0132226	0.8427321	1.218664	0.8904113	1.0000000
1	all_brain_downsampled	0.9227508	0.4931461	1.697124	0.8840815	1.0000000
3	amygdala	2.3695566	1.1874872	4.933618	0.0086392	0.1641453
6	caudate_nucleus	1.0517205	0.6912122	1.596979	0.8389175	1.0000000
7	cerebellum	0.8383851	0.6010500	1.164160	0.2981653	1.0000000
8	globus_pallidus	2.0387026	0.2416209	17.257441	0.4054590	1.0000000
9	hippocampus	1.5846679	0.6920164	3.789543	0.2582880	1.0000000
10	locus_coeruleus	0.7596908	0.3293553	1.793102	0.5477382	1.0000000
11	medial_frontal_gyrus	1.4551317	0.8152096	2.644800	0.2183090	1.0000000
12	medial_temporal_gyrus	1.1838815	0.6344119	2.237078	0.6566256	1.0000000
13	medulla_oblongata	0.9363443	0.4714417	1.941487	0.8669216	1.0000000
14	occipital_cortex	1.3502850	0.7041749	2.619233	0.3576377	1.0000000
15	parietal_gyrus	1.5888767	0.9359603	2.754488	0.0836083	1.0000000
16	putamen	1.3332515	0.8587083	2.108950	0.2118584	1.0000000
17	spinal_cord	0.6135104	0.3116225	1.200928	0.1497080	1.0000000
18	substantia_nigra	0.4415227	0.1416096	1.433082	0.1540947	1.0000000
19	thalamus	0.7381989	0.1648511	4.614510	0.7062522	1.0000000

Plot:



Interpret:

According to the result, the whole brain presented no significant difference in odds ratio, even at a smaller sample size. However, one brain regions, **amygdala**, showed significancy in odds ratio, indicating a more purifying selection in matched promoters than divergent promoters. But the significance could be resulted from the small sample size. Therefore, a larger sample size of promoters or more repeated samples are required.

Discussion

While promoters in most brain regions showed significant purifying selection compared to overall promoters (higher odds ratio towards rare alleles), few present differences between conserved and divergent promoters, which means these pormoters are highly conserved even when the expressing region changed during evolution. However, there are still one brain region, the **amygdala**, that present significantly higher purifying selection in conserved promoters than divergent promoters. However, due to the relative small sample size of rare and common alleles, the result could also be optional. Therefore, a confirmation with more repeated brain samples may help.

Meanwhile, since most brain regions undergo a purifying selection, the change of expression region of the divergent promoters may have undergone a change in their function, while the conserved ones keep their own function. It worth further research to see whether the conserved promoters are functionally important and whether the divergent promoters have loss their function or gain extra function.