PBI – Pokročilá bioinformatika (1. úkol)

Alena Tesařová (xtesar36), říjen 2020

1 Zadání

Pomocí vhodných RNA-seq dat z `http://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP052901` odhadněte expresi jednotlivých genů S100A klastru (mapováním na Vámi připravenou referenci z dané oblasti genomu, převodem na BAM soubor a využitím Vámi připravené anotace klastru BED/GFF3 a samtools s přepínačem -F). Zjistěte jestli data obsahují i některý z intronů (vizuálně/manuálně/IGV, samtools nebo tophat).

2 Postup

2.1 Příprava dat

Jako referenční genom jsem použila genom člověka (GRCh38/hg38), přesněji část úsek: *chr1:153,350,000-153,633,000*, který byl stažen z `https://genome.ucsc.edu/`. Vzorek byl vybrán z `https://trace.ncbi.nlm.nih.gov/` s ID experimentu *SRX856830* a uložen ve formátu *fastq* (obsahuje více informací než fasta).

2.2 Zpracování dat

1. Indexování referenčního genomu pro bowtie2

```
$ bowtie2-build sra_data.fasta human_base
```

2. Vytvoření SAM souboru

```
$ bowtie2 -x indexes/human_base -U experiment_genom/sra_data.fastq -S data_sam.sam
```

3. Vytvoření BAM souboru ze SAM souboru

```
$ samtools view -bS data_sam.sam > data_bam.bam
```

4. Vytvoření setříděného BAM souboru

```
$ samtools sort data_bam.bam -o data_bam.sorted.bam
```

5. Vytvoření indexu SAM souboru (pro IGV)

```
$ samtools index data_bam.sorted.bam
```

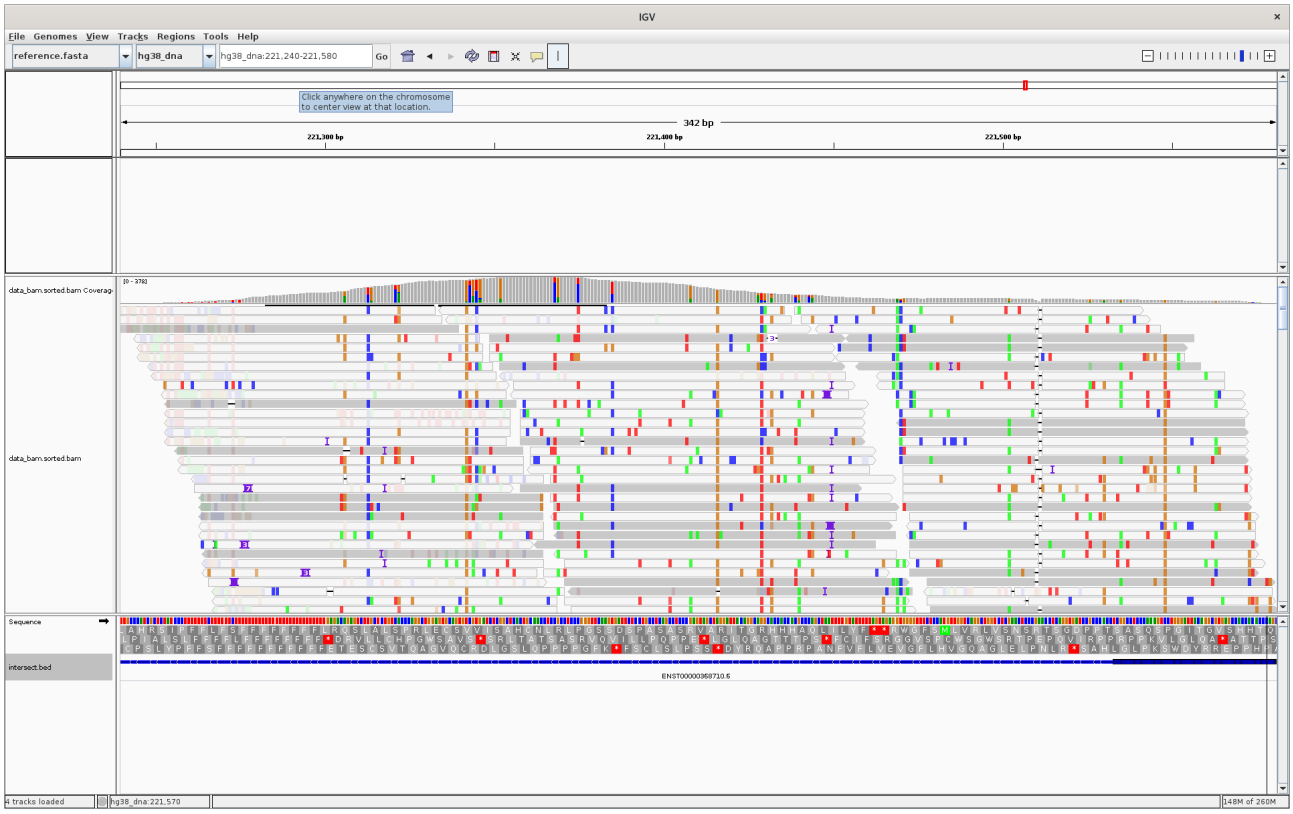
Vytvořený BAM soubor si zobrazíme v prohlížeči IGV.

2.3 Hledání intronů

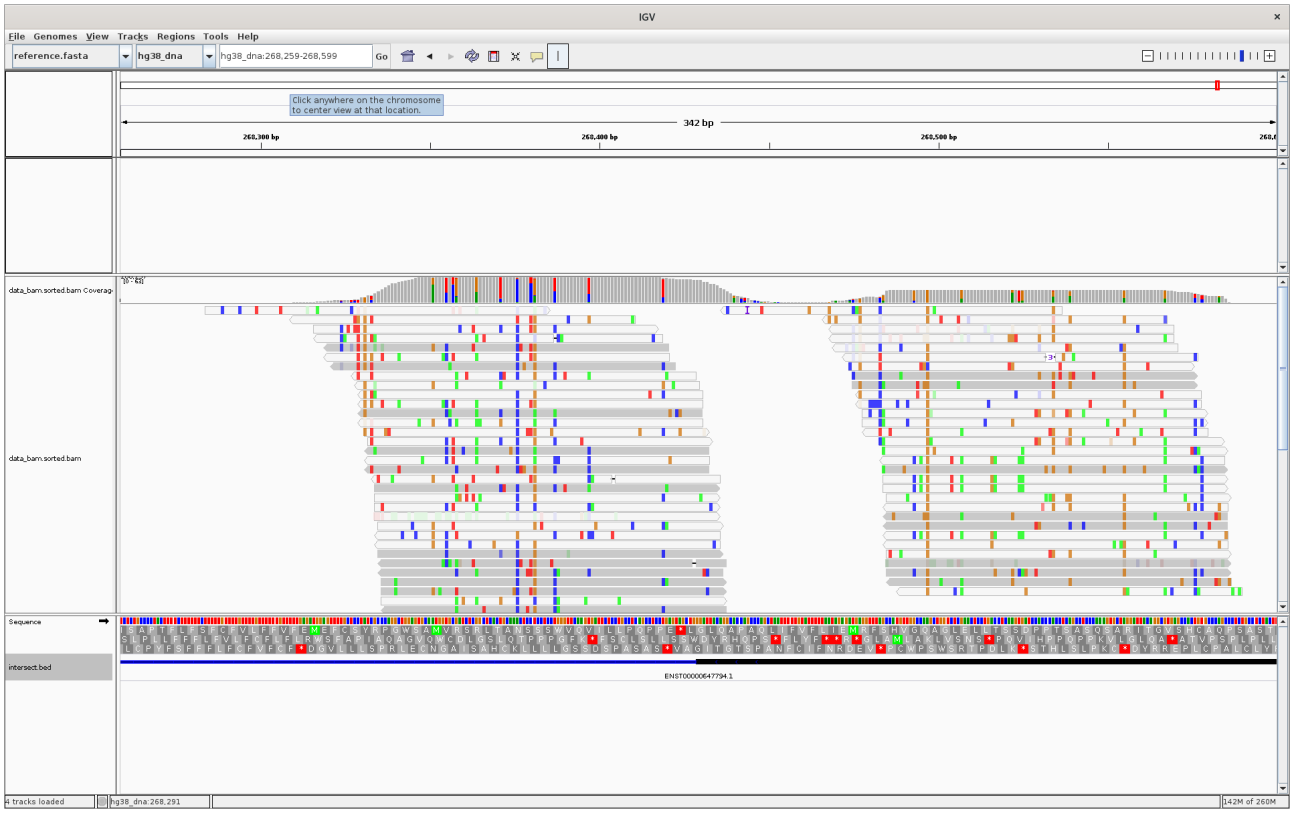
Nejprve je potřeba si stáhnout referenční genom (jeho část) ve formátu BED z genome browseru. Dále je potřeba zjistit, co se namapovalo v oblasti intronů. To zjistíme tak, že si vytvoříme třetí stopu v IGV s anotací genů a porovnáme kolik readů se namapovalo do úseků referenčního fasta souboru (jsou zde vidět i které části jsou introny a exony barevně). Použijeme *intersect*.

```
$ bedtools intersect -sorted -a reference.bed -b data_bam.sorted.bam > intersect.bed
```

Vizuálně v IGV lze hezky vidět 2 místa, kde jsou namapované ready a část jsou introny (modrá barva). První jsem našla kolem pozice 221 300 (Obrázek 1) a druhé kolem pozice 268 350 (Obrázek 2).



Obrázek 1: Přiblížená pozice genomu 221 300 v IGV



Obrázek 2: Přiblížená pozice genomu 268 350 v IGV