Analýza lidského genomu GRCh38

Cílem tohoto projektu je vytvořit skript v jazyce python, který bude analyzovat anotační data lidského genomu (tj. pozice a typy jednotlivých genů, transkriptů, exonů apod.) a vytvoří výstupní tabulku se základními statistikami o přítomnosti jednotlivých typů genů, jejich pokrytí v rámci genomu apod.

Postup práce

1. Stáhněte si soubor *Homo_sapiens.GRCh38.84.gtf.gz* ze stránek genomového prohlížeče Ensembl: ftp://ftp.ensembl.org/pub/release-84/gtf/homo_sapiens/. Tento soubor obsahuje anotaci genů referenčního lidského genomu s označením GRCh38 (z března 2016).

2. Seznamte se s formátem GTF souboru a jeho obsahem. Všimněte si, jakým způsobem jsou uloženy informace o jednotlivých genech. Geny jsou obvykle rozděleny do transkriptů, které reprezentují různé varianty daného genu podléhající transkripci (včetně variant pro alternativní sestřih). Jednotlivé transkripty jsou dále rozděleny do exonů. Pokud se jedná o kódující transkript některého z kódujících genů, potom jsou k dispozici i další informace např. o CDS (Coding segment) oblastech (tj. částech, ze kterých se translací vytváří výsledná sekvence proteinu), UTR (Untranslated Region) oblastech na 5' nebo 3' konci (tj. částech, které nepodléhají translaci, ale jsou potřeba pro uložení signálů pro začátek/konec translace) popřípadě informace o pozicích start nebo stop kodonů.

3. Uložené informace o vybraných genech v GTF souboru si porovnejte s jejich grafickou reprezentací přímo v genomovém prohlížeči. Na hlavní stránce prohlížeče Ensembl zadejte u člověka vyhledávání např. genů HBB (hemoglobin) nebo BRCA2 (breast cancer 2). Zobrazené výsledky porovnejte s informacemi, které jsou uloženy v GTF souboru.

4. Vytvořte skript v jazyce python, který přečte jednotlivé položky z GTF souboru a vytvoří základní statistiku o zastoupení jednotlivých typů genů v lidském genomu (viz příklad výstupní tabulky uvedený níže). Jednotlivé typy genů lze v GTF souboru rozpoznat skrze atribut **gene_biotype.** Podrobnější informace o tom, co jednotlivé gene_biotype znamenají naleznete na stránce http://www.gencodegenes.org/gencode_biotypes.html. Pro zájemce také doporučuji si jednotlivé biotypy dohledat např. na wiki stránkách a dozvědět se o nich více.
   U každého typu genu nás budou dále zajímat následující informace:
   a. Celkový počet genů daného typy v genomu (sloupec Count).
   b. Celkový počet bází, které v souhrnu tyto geny tvoří (sloupec Size [bp]).
   c. Procentuální pokrytí genomu těmito geny. Jinými slovy, kolik procent genomu tyto geny zabírají (sloupec G. Cov. [%]).

5. Informace o jednotlivých typech genů dále rozdělte do čtyř skupin tak, jak uvedeno v příkladu výstupní tabulky:
   a. Kódující geny (biotype = "protein_coding", "IG_*", "TR_*").
   b. Krátké nekódující geny (biotype = "snRNA", "rRNA", "snoRNA", "miRNA", "misc_RNA").
   c. Dlouhé nekódující geny (biotype = "lincRNA", "non_coding", "processed_transcript", "antisense", "3prime_overlapping_ncrna", "sense_intronic", "sense_overlapping", "known_ncrna").
   d. Pseudogeny (biotype = "*pseudogen*").

   K jednotlivým skupinám následně přidejte řádek **Summary** obsahující souhrnné statistiky pro celou skupinu (celkový počet, celkovou velikost v bp, suma pokrytí).

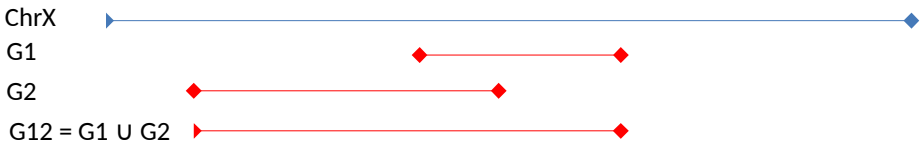
6. Na závěr obdobné statistiky vytvořte pouze pro kódující transkripty a kódující exony (CDS segmenty) tj. pouze pro položky, kde gene_biotype = protein_coding (viz poslední dva řádky výstupní tabulky).
7. Zamyslete se nad získanými výsledky, zejména u kódujících genů porovnejte množství transkribované DNA a kolik z ní ve skutečnosti tvoří protein kódující části.

Doplňující informace k výstupům a skriptu

- Skript bude spouštěn z příkazové řádky a bude mít jeden povinný vstupní parametr reprezentující jméno vstupního GTF souboru.
- Výstupní tabulku bude váš skript generovat na svém standardním výstupu a to ve formátu CSV, kde budete mít jednotlivé položky odděleny středníkem. Tento výstup pak můžete snadno otevřít např. v programu Microsoft Excel nebo OpenOffice Calc.
- Váš skript musí být spustitelný na stroji *merlin.fit.vutbr.cz*.
- Uvedené statistiky počítejte pouze pro geny na chromozomech 1 až 22, X a Y (číslo chromozomu je první položka každého řádku v GTF souboru).
- Součástí GTF souboru nejsou délky jednotlivých chromozomů. Ty lze získat např. na webových stránkách Ensemblu. Pro jednoduchost je zde uvádím (pro chrom. 1 až 22, X, Y):

  chrom_lens = [248956422, 242193529, 198295559, 190214555, 181538259, 170805979, 159345973, 145138636, 138394717, 133797422, 135086622, 133275309, 114364328, 107043718, 101991189, 90338345, 83257441, 80373285, 58617616, 64444167, 46709983, 50818468, 156040895, 57227415]

- Některé geny se mohou v rámci daného gene_biotype překrývat (např. jeden gen může být ukryt uvnitř intronu jiného genu). Potom ale při výpočtu celkového počtu bází (sloupec Size) nebo celkového pokrytí genomu (sloupec G. Cov.) **tyto překrývající se části počítejte pouze jednou!** Na následujícím obrázku je naznačen postup pro výpočet pokrytí dvou překrývajících se elementů na chromozomu ChrX. Stačí pouze zobecnit tento princip na celý genom sloužený z více chromozomů.



Pokrytí chromozomu ChrX elementy G1 a G2 = (délka sloučeného elementu G12/ délka chromozomu) * 100

- Na úrovni jazyka python si můžete např. vytvořit objekty typu:
  - GRange – souřadnice genomového úseku – (chrom, start, stop)
  - GRangeList – seznam objektů typů GRange s operací reduce(), která překrývající se úseky redukuje na nepřekrývající/sloučené úseky.

  Poznámka: Inspirováno knihovnou GenomicRanges jazyka R.

Výstupy projektu

Do informačního systému odevzdejte dva soubory:

- `project.py` – skript v jazyce python.
- `output.csv` – výstup skriptu uložený do souboru.

Příklad výstupní tabulky

Coding genes	**Count**	**Size [bp]**	**G.cov. [%]**
protein_coding			
IG_C_gene			
IG_D_gene			
IG_J_gene			
IG_V_gene			
TR_C_gene			
TR_D_gene			
TR_J_gene			
TR_V_gene			
Summary			
Small non-coding genes	**Count**	**Size [bp]**	**G.cov. [%]**
snRNA			
rRNA			
snoRNA			
miRNA			
misc_RNA			
Summary			
Long non-coding genes	**Count**	**Size [bp]**	**G.cov. [%]**
lincRNA			
non_coding			
processed_transcript			
antisense			
3prime_overlapping_ncrna			
sense_intronic			
sense_overlapping			
known_ncrna			
Summary			
Pseudo genes	**Count**	**Size [bp]**	**G.cov. [%]**
unitary_pseudogene			
IG_C_pseudogene			
translated_processed_pseudogene			
polymorphic_pseudogene			
TR_J_pseudogene			
IG_J_pseudogene			
TR_V_pseudogene			
IG_V_pseudogene			
pseudogene			
unprocessed_pseudogene			
transcribed_unprocessed_pseudogene			
translated_unprocessed_pseudogene			
transcribed_processed_pseudogene			
processed_pseudogene			
transcribed_unitary_pseudogene			
Summary			
Protein Coding Genes	**Count**	**Size [bp]**	**G.cov. [%]**
Coding transcripts			
CDS			