به نام خدا

تحلیل داده‌های حجیم زیستی

نیم‌سال دوم ۹۹-۰۰

مدرس: دکتر مطهری

تمرین سری اول

موعد تحویل: ۲۷ فروردین

مسئله‌ی ۱.(۱۰ نمره) R مقدماتی - بدون استفاده از کتابخانه‌های زیستی

می‌دانیم که در سلول برای ساخت پروتئین، ابتدا رونوشتی از DNA انجام می‌شود و سپس از RNA آن، پروتئین تولید می‌گردد. می‌خواهیم برنامه‌ای بنویسیم که با دریافت یک رشته DNA از ورودی، تمام زیررشته‌هایی که از آن‌ها پروتئین ساخته می‌شود را تولید نماید. (توجه: رشته DNA داده شده از نقاط اگزونی گرفته شده است.)

input	*output*
AGCCATGTAGCTAACTCAGGTTACATGGGGATGACCCCGCGAC TTGGATTAGAGTCTCTTTTGGAATAAGCCTGAATGATCCGAGTA GCATCTCAG	MLLGSFRLIPKETLIQVAGSSPCNLS M MGMTPRLGLESLLE MTPRLGLESLLE

مسئله‌ی ۲.(۳۰نمره) استفاده از کتابخانه‌های زیستی و کتابخانه‌های موجود در R

در این سایت UCSC Genome Browser ابزاری به نام Genome Browser وجود دارد که با انتخاب گونه‌ی مد نظر و جستجوی سمبل یک ژن، محل قرار گرفتن آن ژن در کروموزوم و سایر اطلاعات مربوط به آن را بیابیم.

الف) (**۲ نمره**) با استفاده از سایت گفته شده، سمبل RPL5 را در ژنوم انسان جستجو نمایید.

ب) (**٦ نمره**)کتابخانه‌هایی که در این سوال می‌خواهیم بررسی کنیم BSgenome، org.Hs.eg.db، TxDb است. درمورد هرکدام توضیح کوتاهی دهید.

ج) (**۲۲نمره**) می‌خواهیم کاری مشابه Genome Browser انجام دهیم. برای همین برنامه‌ای بنویسید که ابتدا سمبل یک ژن از انسان (hg19) را دریافت نماید و با استفاده از کتابخانه‌های زیستی، تمامی رونوشت‌های ژن مربوطه استخراج نماید. ابتدا در خروجی کنسول R ، نسبت طول ژن را به طول کروموزومی که در آن قرار دارد ، محاسبه کند. سپس بخش های exon و intron آن را مشخص نموده و در نهایت با استفاده از کتابخانه‌ی genemodel آن را رسم نماید.

مسئله‌ی ۳.(۶۰نمره) کارگاه NGS

یک فولدر به نام Q3 بسازید. کروموزوم شماره‌ی ۱۹ انسان را از این آدرس دانلود کنید. توجه داشته‌باشید که در این تمرین، از این کروموزوم به عنوان مرجع استفاده می‌کنیم. بنابراین برای انجام این تمرین، نیازی به دانلود کل ژنوم انسان ندارید.

۱) (۲۵نمره) ساخت نمونه

از کروموزوم ۱۹، بخشی با طول ۵ میلیون base را به صورت دلخواه انتخاب کرده و در یک فایل با فرمت fasta و با نام partof19.fa ذخیره کنید. ۱۰۰۰جهش با شرایط زیر را با استفاده از یک کد روی این فایل ایجاد کنید:

۱ – ۴۰۰ جهش از نوع SNV

۲- ۳۰۰ جهش از نوع insertion

۳- ۳۰۰ جهش از نوع deletion

طول ناحیه Insertion یا deletion را متغیر بین ۱ تا ۵ base در نظر بگیرید. موقعیت تمامی جهش‌ها را کاملاً تصادفی و با توزیع یکسان در سراسر ژنوم انتخاب کنید.

ژنوم دارای جهش را با نام Sample.fa و در آدرس HW1/Q3/Part1/Sample.fa ذخیره کنید.

همچنین یک فایل به نام Sample.vcf بسازید و تمامی جهش‌هایی که ایجاد کرده‌اید را در این فایل ذخیره کنید. توجه کنید که این فایل را باید خودتان بسازید و نیازی به ابزارهای دیگر ندارید. تنها کافیست اطلاعات مربوط به جهش‌هایی که ایجاد کرده‌اید را در یک فایل که شامل چهار ستون زیر است، ذخیره کنید (بدیهی‌ست که در ستون اول، تنها نام chr19 باید باشد. همچنین pos باید موقعیت جهش نسبت به ابتدای کروموزوم ۱۹ را مشخص کند).:

CHROM POS REF ALT

Sample.vcf را نیز در آدرس HW1/Q3/Part1/Sample.vcf قرار دهید.

کدهای مربوط به این بخش را نیز در این آدرس قرار دهید:

HW1/Q3/Part1/Code

۲) (۱۰نمره) شبیه سازی

یک ابزار مناسب جهت شبیه سازی توالی های Whole Genome Sequencingپیدا کنید. معیار شما برای انتخاب این ابزار شامل قابلیت های آن، میزان استفاده از آن توسط دیگر پژوهشگران، تعداد ارجاع های انجام شده روی مقاله مربوط به آن (متناسب با سال انتشار آن)، به روز بودن و پشتیبانی از آن و داشتن مستندات کافی خواهد بود.
برای این کار می توانید از مقالاتی که ابزارهای مختلف را مقایسه می‌کنند هم استفاده کنید.

ابزارهایی که در فهرست کاندیداهای شما بوده‌اند و دلایل انتخاب ابزار نهایی خود را در یک فایل توضیح داده و در این آدرس قرار دهید:

HW1/Q3/Part2/Simulator.pdf

با استفاده از ابزاری که انتخاب کرده‌اید، یک جفت فایل fastq شامل توالی هایی به طول 100حرف به عنوان Paired-end Whole Genome Sequencing از ژنوم نمونه ساخته شده در مرحله قبل، با coverage برابر با 30x بسازید. و نتیجه را در آدرس‌های زیر ذخیره کنید:

HW1/Q3/Part2/Sample_1.fastq
HW1/Q3/Part2/Sample_2.fastq

همچنین تمام کدها و اسکریپت های خود را برای این قسمت در شاخه زیر ذخیره کنید:

HW1/Q3/Part2/Code

۳) (۲۵نمره) مقایسه

فایل‌های fastq ساخته شده در مرحله‌ی قبل را مطابق آنچه در ویدئو توضیح داده‌شد، آنالیز کنید. توجه داشته باشید، همان طور که در ابتدای تمرین توضیح داده‌شد، در این تمرین از فایل chr19.fa به عنوان ژنوم مرجع استفاده می‌شود. بنابراین قبل از شروع alignment برای این فایل با استفاده از ابزار picard، دیکشنری، و با استفاده از bwa ، فایل index بسازید.

علاوه بر ابزار GATK ، جهش‌های ژنومی را با استفاده از دو ابزار Samtools و FreeBayes نیز استخراج کنید و فایل vcf حاصل از هر یک را در آدرس‌های زیر قرار دهید:

HW1/Q3/Part3/GATK4.vcf
HW1/Q3/Part3/Samtools.vcf
HW1/Q3/Part3/FreeBayes.vcf

حال این ۳ فایل را با هم و هم‌چنین با فایل vcf ای که در بخش اول ساختید مقایسه کنید نتیجه را در یک گزارش توضیح دهید. کدام یک از این سه ابزار نتیجه‌ی بهتری تولید کرده‌است؟

کدها و script های مربوط به این بخش را نیز در آدرس زیر ذخیره کنید:

HW1/Q3/Part3/Code

گزارش مربوط به این تمرین را نیز در فولدر Ex ذخیره کنید. این گزارش باید شامل موارد زیر باشد:

۱- توضیحات مربوط به کدهای تمامی بخش‌ها

۲- تمامی کدهای استفاده شده در هر مرحله در خلال توضیحات گزارش برای هر قسمت

۳- نتیجه‌گیری بخش آخر از مقایسه‌ی ۳ ابزار معرفی شده.

فولدر نتایج تمرین خود را به صورت یک فایل فشرده بارگذاری کنید.