
بسم الله الرحمن الرحيم

وزارت علوم، تحقیقات و فناوری
دانشگاه تربت حیدریه

ژنومیک محاسباتی با R

Population Genomics with R

آلتونا آکالین، وردا فرانک، بورا اویار و جاناتان رونن

مسعود علی پناه

(دانشیار، گروه تولیدات گیاهی، دانشگاه تربت حیدریه)

ایمان یوسفی‌جوان

(استادیار، گروه تولیدات گیاهی، دانشگاه تربت حیدریه)

(۱۴۰۴)

عنوان و نام پدیدآور	:	ژنومیک محاسباتی با R/نویسندگان آلتونا آکالین...[و دیگران] ؛ مترجمان مسعود علی‌پناه، ایمان یوسفی‌جوان.
مشخصات نشر	:	تربت حیدریه: دانشگاه تربت حیدریه، انتشارات، سال ۱۴۰۳
مشخصات ظاهری	:	ق،۷۱۲ص.: مصور،نمودار
شابک	:	۹۷۸-۶۰۰-۸۳۳۵-۴۰-۵
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	عنوان اصلی:.Computational Genomics with R, 2020
یادداشت		نویسندگان آلتونا آکالین، وردا فرانک، بورا اویار،جاناتان رونن.
یادداشت		کتابنامه:ص.۶۹۴-۷۱۱.
موضوع	:	ژنتیک پزشکی Medical genetics آر (زبان برنامه‌نویسی کامپیوتر) R (Computer program language) ژنتیک -- روش‌های آماری Genetics -- Statistical methods
شناسه افزوده	:	آکالین، آلتونا
شناسه افزوده	:	Akalin, Altuna
شناسه افزوده	:	علی‌پناه، مسعود، ۱۳۴۸ -، مترجم
شناسه افزوده		یوسفی جوان، ایمان، مترجم
شناسه افزوده		دانشگاه تربت حیدریه
رده بندی کنگره	:	RB۱۵۵
رده بندی دیویی	:	۰۴۲/۶۱۶
شماره کتابشناسی ملی	:	۹۸۲۰۳۱۳

این اثر مشمول قانون حمایت از مؤلفان و مصنفان و هنرمندان است. هرکس تمام یا قسمتی از این اثر را بدون اجازه ناشر، نشر، پخش یا عرضه کند مورد پیگرد قانونی قرار خواهد گرفت.

ژنومیک محاسباتی با R

نویسنده: آلتونا آکالین، وردا فرانک، بورا اویار،جاناتان رونن.

مترجم: مسعود علی‌پناه، ایمان یوسفی‌جوان.

چاپ اول

بها: تومان

نشانی ناشر: تربت حیدریه، کیلومتر هفت جاده مشهد، دانشگاه دولتی تربت حیدریه

مسئولیت کلیه مطالب این کتاب به عهده نگارنده می‌باشد. دانشگاه تربت حیدریه هیچگونه مسئولیتی در قبال صحت و سقم مطالب ندارد.

عنوان مطالب **صفحه**

فصل ۳

فصل ٤

فصل ۵

فصل ٦

فصل ۷

فصل ۸

فصل ۹

فصل ۱۰

فصل ۱۱

پیشگفتار

هدف این کتاب ارائه مبانی تجزیه‌وتحلیل داده‌ها برای ژنومیک است. ما این کتاب را بر اساس دوره‌های ژنومی محاسباتی که هر سال ارائه می‌دهیم، تهیه کرده‌ایم. ما همواره مخاطبان بین‌رشته‌ای با پیشینه‌ای از فیزیک، زیست‌شناسی، پزشکی، ریاضی، علوم کامپیوتر یا سایر زمینه‌های کمی داشته‌ایم. ما می‌خواهیم این کتاب نقطه شروعی برای دانشجویان ژنومی محاسباتی و راهنمایی برای تجزیه‌وتحلیل بیشتر داده‌ها در موضوعات خاص‌تر در ژنومیک باشد. به همین دلیل است که ما سعی کردیم موضوعات مختلفی از برنامه‌نویسی گرفته تا زیست‌شناسی ژنوم پایه را پوشش دهیم. ازآنجایی‌که این رشته بین‌رشته‌ای است، برای افراد با پیشینه‌های مختلف به نقاط شروع متفاوتی نیاز دارد. یک زیست‌شناس ممکن است بخش‌های مربوط به زیست‌شناسی ژنوم اولیه را نادیده بگیرد و با برنامه‌نویسی R شروع کند، درحالی‌که یک دانشمند کامپیوتر ممکن است بخواهد با زیست‌شناسی ژنوم شروع کند. به همین ترتیب، یک فرد باتجربه‌تر ممکن است بخواهد در صورت نیاز به انجام نوع خاصی از تحلیل، اما بدون تجربه قبلی، به این کتاب مراجعه کند.

نسخه آنلاین این کتاب تحت مجوز Creative Commons Attribution-NonCommercial-ShareAlike ۴/۰ مجوز بین‌المللی[1] دارد.

این کتاب برای چه کسانی است؟

این کتاب شامل جنبه‌های عملی و نظری ژنومیک محاسباتی است. زیست‌شناسی و پزشکی بیش از هر زمان دیگری داده تولید می‌کنند؛ بنابراین، ما باید افراد بیشتری را بامهارت‌های تجزیه‌وتحلیل داده‌ها و درک ژنومیک محاسباتی آموزش دهیم. ازآنجایی‌که ژنومیک محاسباتی بین‌رشته‌ای است، هدف این کتاب برای زیست‌شناسان، دانشمندان علوم پزشکی،

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/

دانشمندان کامپیوتر و افرادی با سایر زمینه‌های کمی است. ما این کتاب را برای مخاطبان زیر نوشتیم:

- زیست‌شناسان و دانشمندان پزشکی که داده‌ها را تولید می‌کنند و خود مشتاق تجزیه‌وتحلیل آن هستند.
- دانش‌آموزان و محققانی که به طور رسمی شروع به تحقیق در مورد ژنومیک محاسباتی می‌کنند یا از آن استفاده می‌کنند، دانش گسترده‌ای در حوزه خاص ندارند، اما حداقل درک سطح مبتدی در زمینه‌های کمی به‌عنوان‌مثال، ریاضی، آمار، دارند.
- محققان باتجربه به دنبال دستورالعمل‌ها یا روش‌های سریع برای شروع انجام تجزیه‌وتحلیل داده‌های خاص مربوط به ژنومیک محاسباتی هستند.

خروجی چه خواهد بود؟

این منبع مهارت‌ها را توصیف می‌کند و روش‌هایی را ارائه می‌دهد که به خوانندگان کمک می‌کند تا داده‌های ژنومیک خود را تجزیه‌وتحلیل کنند.

بعد از خواندن:

- اگر با R آشنا نیستید، اصول اولیه R را دریافت کرده و مستقیماً به سراغ استفاده‌های تخصصی از R برای ژنومیک محاسباتی خواهید رفت.
- فواصل ژنومی و عملیات روی آن‌ها مانند همپوشانی را درک خواهید کرد.
- شما می‌توانید از R و کتابخانه گسترده بسته‌های نرم‌افزاری آن برای انجام تجزیه‌وتحلیل توالی، مانند محاسبه محتوای GC برای بخش‌های معینی از ژنوم یا یافتن مکان‌های اتصال فاکتور رونویسی استفاده کنید.
- شما با تکنیک‌های تصویرسازی مورداستفاده در ژنومیک، مانند نقشه‌های حرارتی، نمودارهای متا ژن و تصویرسازی مسیر ژنومی آشنا خواهید شد.

- شما با تکنیک‌های یادگیری نظارت شده و بدون نظارت که در مدل‌سازی داده‌ها و تجزیه‌وتحلیل اکتشافی داده‌های با ابعاد بالا مهم هستند آشنا خواهید شد.
- شما با تجزیه‌وتحلیل مجموعه‌داده‌های مختلف توالی‌یابی با کارایی بالا (RNA-seq، ChiP-seq، BS- seq و ادغام چندگانه امیکس) که عمدتاً از ابزارهای مبتنی بر R استفاده می‌کنند، آشنا خواهید شد.

ساختار کتاب

این کتاب با این ایده طراحی شده است که درک عملی و مفهومی روش‌های تجزیه‌وتحلیل داده‌ها به همان اندازه، اگر نه مهم‌تر از درک نظری، مانند استخراج دقیق معادلات در آمار یا یادگیری ماشین مهم است. به همین دلیل است که ابتدا سعی می‌کنیم توضیحی مفهومی از مفاهیم، ارائه دهیم سپس سعی می‌کنیم بخش‌های اساسی فرمول‌های ریاضی را برای درک دقیق‌تر ارائه دهیم. در این مفهوم، ما همیشه کد یک کار تجزیه‌وتحلیل داده خاص را نشان داده و توضیح می‌دهیم. ما همچنین به خوانندگانی که مایل‌اند درک نظری عمیق‌تری از روش‌ها یا مفاهیم مرتبط با تجزیه‌وتحلیل داده‌ها به دست آورند، منابع اضافی مانند کتاب‌ها، وب‌سایت‌ها، سخنرانی‌های ویدئویی و مقالات علمی ارائه می‌کنیم.

فصل ۱: "مقدمه‌ای بر ژنومیکس" به معرفی مفاهیم اساسی در زیست‌شناسی ژنوم و ژنومیک می‌پردازد. درک این مفاهیم برای ژنومیک محاسباتی مهم است.

فصل ۲: "مقدمه‌ای بر R برای تجزیه‌وتحلیل داده‌های ژنومی" علاوه بر پارادایم‌های رایج تجزیه‌وتحلیل داده‌ها که در تجزیه‌وتحلیل داده‌های ژنومی مشاهده می‌کنیم، مهارت‌های پایه‌ای لازم R برای دنبال‌کردن کتاب را فراهم می‌کند. فصل ۳: "آمار برای ژنومیکس"، فصل ۴: "تجزیه‌وتحلیل داده‌های اکتشافی با یادگیری ماشین بدون ناظر" و فصل ۵: "مدل‌سازی پیش‌بینی با یادگیری ماشینی با ناظر" مهارت‌های کمی لازم را که فرد هنگام تجزیه‌وتحلیل داده‌های ژنومیک با ابعاد بالا به آن نیاز دارد، معرفی می‌کند.

فصل ۶: "عملیات بر روی فواصل ژنومی و محاسبات ژنومی" ابزارهای اساسی در مواجه با فواصل ژنومی و ارتباط آن‌ها با یکدیگر بر روی ژنوم را معرفی می‌کند. علاوه بر این، این فصل انواع روش‌های تصویرسازی داده‌های ژنومی را معرفی می‌کند. مهارت‌های معرفی‌شده در این فصل، مهارت‌های کلیدی هستند که برای کار با داده‌های ژنومی پردازش‌شده که از طریق پایگاه‌های عمومی داده قابل‌دسترس هستند مانند Ensembl و مرورگر UCSC، موردنیاز هستند.

فصل‌های بعدی به تجزیه‌وتحلیل خاص داده‌های توالی‌یابی با توان بالا و ادغام انواع مختلف مجموعه‌داده‌ها می‌پردازد. فصل ۷: "بررسی کیفیت، پردازش و ترازسازی خوانش‌های توالی با توان بالا" بررسی‌های کیفیتی را که باید در مورد خواندن توالی انجام شود و روش‌های مختلف برای پردازش بیشتر آن‌ها معرفی می‌کند. در فصل‌های ۸، ۹ و ۱۰ به تجزیه‌وتحلیل RNA-seq، تجزیه‌وتحلیل ChiP-seq و تجزیه‌وتحلیل BS-seq پرداخته می‌شود. فصل آخر، فصل ۱۱: "تحلیل چندگانه امیکس" به روش‌هایی برای ادغام مجموعه‌داده‌های امیکس متعدد می‌پردازد.

اکثر فصل‌ها دارای تمرین‌هایی هستند که برخی از نکات مهم معرفی شده در فصل‌ها را تقویت می‌کند. تمرینات در دسته‌های مبتدی، متوسط و پیشرفته طبقه‌بندی می‌شوند. اگر در یک موضوع خاص به‌خوبی آشنا هستید، ممکن است بخواهید از تمرینات سطح مبتدی صرف‌نظر کنید.

به طور خلاصه، این کتاب یک راهنمای جامع برای ژنومیک محاسباتی است. برخی از بخش‌ها به‌خاطر مخاطبان بین‌رشته‌ای گسترده و تکمیل اطلاعات آن‌ها وجود دارد و همه بخش‌ها به یک اندازه برای همه خوانندگان این مخاطب گسترده مفید نیستند.

اطلاعات نرم‌افزار و قراردادها

نام بسته‌ها و کد درون‌خطی و نام فایل‌ها با فونت ماشین‌تحریر (به‌عنوان‌مثال methylKit) قالب‌بندی می‌شوند. نام توابع با پرانتز دنبال می‌شود (به‌عنوان‌مثال genomation: ScoreMatrix ( )). عملگر دو نقطه‌ای: به معنای دسترسی به یک شیء از یک بسته است.

کنوانسیون عملگر انتساب

به طور سنتی، <- عملگر انتساب ترجیحی است. بااین‌حال، در سراسر کتاب ما از = و <- به‌عنوان عملگر انتساب به‌جای یکدیگر استفاده می‌کنیم.

بسته‌های موردنیاز برای اجرای کد کتاب

این کتاب در درجه اول در مورداستفاده از بسته‌های Rبرای تجزیه‌وتحلیل داده‌های ژنومیک است، بنابراین اگر می‌خواهید تجزیه‌وتحلیل را در این کتاب تکرار کنید، باید بسته‌های مربوطه را در هر فصل با استفاده از install.packages یا BiocManager::install نصب کنید. در هر فصل، هنگامی که از توابع موردنیاز از بسته‌های مربوطه استفاده می‌کنیم، بسته‌های لازم را با تابع ()library یا ()require بارگذاری می‌کنیم. با نگاه‌کردن به فراخوآن‌ها، می‌توانید ببینید چه بسته‌هایی برای کد آن قطعه یا فصل موردنیاز است. اگر نیاز به نصب تمام بسته‌های الحاقی برای کتاب دارید، می‌توانید دستور زیر را اجرا کنید و در حین انتظار یک فنجان چای بنوشید.

```
if (requireNamespace("BiocManager", quietly = TRUE))
   install.packages("BiocManager")
BiocManager::installc\('qvalue','plot3D','ggplot2','pheatmap','cowplot',
                  'cluster', 'NbClust', 'fastICA', 'NMF','matrixStats',
                  'Rtsne', 'mosaic', 'knitr', 'genomation',
                  'ggbio', 'Gviz', 'DESeq2', 'RUVSeq',
                  'gProfileR', 'ggfortify', 'corrplot',
                  'gage', 'EDASeq', 'citr', 'formatR',
                  'svglite', 'Rqc', 'ShortRead', 'QuasR',
                  'methylKit','FactoMineR', 'iClusterPlus',
                  'enrichR','caret','xgboost','glmnet',
                  'DALEX','kernlab','pROC','nnet','RANN',
                  'ranger','GenomeInfoDb', 'GenomicRanges',
                  'GenomicAlignments', 'ComplexHeatmap', 'circlize',
                  'rtracklayer', 'BSgenome.Hsapiens.UCSC.hg38',
                  'BSgenome.Hsapiens.UCSC.hg19','tidyr',
                  'AnnotationHub', 'GenomicFeatures', 'normr',
                  'MotifDb', 'TFBSTools', 'rGADEM', 'JASPAR2018'
```

داده‌های مورداستفاده در این کتاب

ما در سراسر کتاب به داده‌های بسته‌های مختلف R و Bioconductor تکیه می‌کنیم. برای مجموعه‌داده‌هایی که با آن بسته‌ها ارسال نمی‌شوند، ما بسته compGenomRData[2] خودمان را ایجاد کردیم. می‌توانید این بسته را از طریق devtools::install_github ("compgenomr/compGenomRData") نصب کنید. ما از تابع system.file() برای دریافت مسیر فایل‌ها استفاده می‌کنیم. ما متوجه شدیم که بسیاری از کاربران بی‌تجربه در مورد این عملکرد گیج شده‌اند. این تابع فقط خروجی‌های مسیر کامل فایل نصب شده با بسته داده را می‌دهد.

تمرین‌های موجود در کتاب

در پایان هر فصل مجموعه‌ای از تمرین‌ها وجود دارد. تمرین‌ها در بخش‌های موضوعی که از بخش‌های اصلی فصل پیروی می‌کنند، جدا شده‌اند. علاوه بر این، هر تمرین بر اساس سختی آن به‌عنوان "مبتدی"، "متوسط" و "پیشرفته" طبقه‌بندی می‌شود. تمرین‌های سطح مبتدی را معمولاً می‌توان با تغییردادن کد موجود در فصل انجام داد. تمرین‌های سطح پیشرفته معمولاً به ترکیبی از کدها از بخش‌ها یا فصل‌های مختلف نیاز دارند. سطح متوسط جایی دراین‌بین است. راه‌حل‌های تمرینات در آدرس زیر موجود است.

https://github.com/compgenomr/exercises

حرف‌های تکراری

این کتاب با R 4.0.0 و بسته‌های زیر گردآوری شده است. ما فقط بسته‌های اصلی و نسخه‌های آن‌ها را و نه ملحقات آن را فهرست کرده‌ایم.

```
## qvalue_2.20.0 | plot3D_1.3 | ggplot2_3.3.1 | pheatmap_1.0.12
## cowplot_1.0.0 | cluster_2.1.0 | NbClust_3.0 | fastICA_1.2.2
## NMF_0.23.0 | matrixStats_0.56.0 | Rtsne_0.15 | mosaic_1.7.0
## knitr_1.28 | genomation_1.20.0 | ggbio_1.36.0 | Gviz_1.32.0
```

[2] https://github.com/compgenomr/compGenomRData

DESeq2_1.28.1 | RUVSeq_1.22.0 | gProfileR_0.7.0 | ggfortify_0.4.10
corrplot_0.84 | gage_2.37.0 | EDASeq_2.22.0 | citr_0.3.2
formatR_1.7 | svglite_1.2.3 | Rqc_1.22.0 | ShortRead_1.46.0
QuasR_1.28.0 | methylKit_1.14.2 | FactoMineR_2.3 |
iClusterPlus_1.24.0
enrichR_2.1 | caret_6.0.86 | xgboost_1.0.0.2 | glmnet_4.0
DALEX_1.2.1 | kernlab_0.9.29 | pROC_1.16.2 | nnet_7.3.14
RANN_2.6.1 | ranger_0.12.1 | GenomeInfoDb_1.24.0 |
GenomicRanges_1.40.0
GenomicAlignments_1.24.0 | ComplexHeatmap_2.4.2 | circlize_0.4.9 |
rtracklayer_1.48.0
tidyr_1.1.0 | AnnotationHub_2.20.0 | GenomicFeatures_1.40.0 |
normr_1.14.0
MotifDb_1.30.0 | TFBSTools_1.26.0 | rGADEM_2.36.0 |
JASPAR2018_1.1.1
BSgenome.Hsapiens.UCSC.hg38_1.4.3 |
BSgenome.Hsapiens.UCSC.hg19_1.4.3

سپاسگزاری‌ها

مایلم از انجمن‌های R و Bioconductor برای توسعه و نگهداری کتابخانه‌ها برای تجزیه‌وتحلیل داده‌های ژنومی تشکر کنم. بدون تلاش و فداکاری مستمر آن‌ها، نوشتن چنین کتابی ممکن نخواهد بود.

همچنین از تمامی مربیان، همکاران و کارفرمایان گذشته و حال خود تشکر می‌کنم. تعامل با آن‌ها انگیزه نوشتن چنین کتابی و سازماندهی و تدریس دوره‌های عملی ژنومیک محاسباتی را فراهم کرد.

می‌خواهم از جان کیمل، ویراستار Chapman & Hall/CRC که در انتشار این کتاب به من کمک کرد، تشکر کنم. کار با او لذت‌بخش بود. او سخاوتمندانه موافقت کرد که به من اجازه دهد نسخه آنلاین این کتاب را نگه دارم، تا بتوانم پس از چاپ آن را به‌روز کنم. این یک سفر طولانی برای من بوده است. نوشتن بخش‌هایی از این کتاب را از اوایل سال ۲۰۱۳ شروع کردم. اگر ودران فرانکه، بورا اویار و جاناتان رونن نبودند، حتی بیشتر طول می‌کشید. آن‌ها با مهربانی موافقت کردند که در نگارش فصل‌های گمشده مشارکت کنند و کار بزرگی انجام دادند. من از کمک‌های آن‌ها سپاسگزارم.

افراد زیر با مهربانی برای رفع اشتباهات تایپی و کد و پیشنهادهای مختلف کمک کردند: توماس شالچ، الکس گوسدشان، رودریگو اوگاوا، فی ژائو، جاناتان کیت، جانانی راوی، کریستین شودوما، ساموئل اسلدزیسکی، دانیا هامو و سروش نیکومب.

آلتونا آکالین

برلین، آلمان

درباره نویسندگان

دکتر آلتونا آکالین[3] ساختار کتاب را سازماندهی کرد، بیشتر کتاب را نوشت و بقیه را ویرایش کرد. او یک دانشمند بیوانفورماتیک و رئیس بیوانفورماتیک و پلتفرم علوم داده‌های امیکس در مؤسسه زیست‌شناسی سیستم‌های پزشکی برلین، مرکز Max Delbrück در برلین است. او از سال ۲۰۰۲ درحال‌توسعه روش‌های محاسباتی برای تجزیه‌وتحلیل و یکپارچه‌سازی مجموعه‌های داده‌های ژنومیک در مقیاس بزرگ است. او علاقه‌مند به استفاده از یادگیری ماشینی و آمار برای کشف الگوهای مرتبط با متغیرهای مهم بیولوژیکی مانند وضعیت و نوع بیماری است. او در ایالات متحده آمریکا، نروژ، ترکیه، ژاپن و سوئیس زندگی کرد تا کارهای تحقیقاتی و آموزش مرتبط با ژنومیک محاسباتی را دنبال کند. هدف اساسی کار فعلی او استفاده از امضاهای مولکولی پیچیده برای ارائه سیستم‌های پشتیبانی تصمیم برای تشخیص بیماری و کشف نشانگرهای زیستی است. وی علاوه بر تلاش‌های پژوهشی و مدیریت یک آزمایشگاه علمی، از سال ۲۰۱۵، در دوره‌های ژنومی محاسباتی در برلین با شرکت‌کنندگانی از سراسر جهان، سازماندهی و تدریس می‌کند. این کتاب بیشتر نتیجه مطالبی است که برای آن‌ها و تلاش‌های قبلی تدریس در کالج پزشکی ویل کورنل در نیویورک و مؤسسه فردریش میشر در بازل، سوئیس تهیه شده است.

دکتر آکالین و نویسندگان مشارکت‌کننده زیر چندین دهه تجربه ترکیبی در تجزیه‌وتحلیل داده‌ها برای ژنومیک دارند. آن‌ها توسعه‌دهندگان بسته‌های Bioconductor مانند methylKit[4]، genomation[5]، RCAS[6] و netSmooth[7] هستند. علاوه بر این، آن‌ها نقش کلیدی در توسعه تمام‌وکمال مراحل تجزیه‌وتحلیل داده‌های ژنومیک برای RNA-seq، ChiP-seq، Bisulfite-seq و RNA-seq تک‌سلولی به نام PiGx[8] ایفا کرده‌اند.

---

[3] https://github.com/al2na
[4] https://bioconductor.org/packages/release/bioc/html/methylKit.html
[5] https://bioconductor.org/packages/release/bioc/html/genomation.html
[6] https://bioconductor.org/packages/release/bioc/html/RCAS.html
[7] https://bioconductor.org/packages/release/bioc/html/netSmooth.html
[8] http://bioinformatics.mdc-berlin.de/pigx/

نویسندگان مشارکت‌کننده

دکتر بورا اویار[9] در فصل ۸، "تحلیل RNA-seq" مشارکت داشت. او آموزش بیوانفورماتیک خود را در دانشگاه سابانچی (استانبول/ترکیه) آغاز کرد و از آنجا مدرک کارشناسی خود را گرفت. بعداً از دانشگاه سایمون فریزر (ونکوور/کانادا) مدرک کارشناسی‌ارشد گرفت و سپس از آزمایشگاه بیولوژی مولکولی اروپا در هایدلبرگ/آلمان مدرک دکترا گرفت. از سال ۲۰۱۵، او به‌عنوان دانشمند بیوانفورماتیک در پلتفرم بیوانفورماتیک و پلتفرم علوم داده Omics در مؤسسه زیست‌شناسی سیستم‌های پزشکی برلین مشغول به کار است. او از طریق تحقیق، همکاری، خدمات و توسعه روش تجزیه‌وتحلیل داده‌ها به پلت فرم بیوانفورماتیک کمک کرده است. علاقه اصلی تحقیقاتی فعلی او ادغام انواع مختلف مجموعه‌داده‌های Omics برای کشف نشانگرهای زیستی پیش‌آگهی/تشخیصی سرطان‌ها است.

دکتر ودران فرانکه[10] در فصل ۹، "تحلیل ChIP-seq" مشارکت کرد. وی دکترای خود را از دانشگاه زاگرب دریافت کرد. کار او بر روی بیو ژنز و عملکرد مولکول‌های RNA کوچک در طول جنین‌زایی اولیه و ایجاد پرتوانی متمرکز بود. قبل از دکترا، او به‌عنوان محقق علمی زیر نظر بوریس لنهارد در دانشگاه برگن، نروژ، با تمرکز بر اصول عملکردهای تقویت‌کننده ژن کار می‌کرد. او تحقیقات خود را در بستر علم داده‌های بیوانفورماتیک و Omics در مؤسسه زیست‌شناسی سیستم پزشکی برلین ادامه می‌دهد. او ابزارهایی را برای یکپارچه‌سازی داده‌های چندگانه Omics ، با تمرکز بر توالی‌یابی RNA تک‌سلولی، و اپی ژنومیک توسعه می‌دهد. دانش یکپارچه او از فیزیولوژی سلولی همراه بامهارت او در تجزیه‌وتحلیل داده‌ها او را قادر می‌سازد تا راه‌حل‌های خلاقآن‌های برای مشکلات زیستی دشوار بیابد.

---

[9] https://bioconductor.org/packages/release/bioc/html/methylKit.html
[10] https://github.com/frenkiboy

دکتر *جاناتان رونن*[11] در فصل ۱۱، "تحلیل چندگانه امیکس" مشارکت داشت. دکتر رونن مدرک کارشناسی‌ارشد خود را در رشته مهندسی کنترل از دانشگاه علم و صنعت نروژ در سال ۲۰۱۰ دریافت کرد. سپس به‌عنوان توسعه‌دهنده نرم‌افزار در اسلو، بروکسل و مونیخ مشغول به کار شد. در آن زمان، او همچنین در تیم مؤسس www.holderdeord.no بود، وب‌سایتی که رای‌های پارلمان نروژ را به وعده‌های داده شده در اعلامیه‌های حزبی پیوند می‌دهد. در سال‌های ۲۰۱۴–۲۰۱۵، او به‌عنوان دانشمند داده در آزمایشگاه رسانه‌های اجتماعی و مشارکت سیاسی دانشگاه نیویورک کارکرد. در آن زمان، او همچنین www.lahadam.co.il را راه‌اندازی کرد، وب‌سایتی که پست‌های فیس‌بوک سیاستمداران اسرائیلی را ردیابی می‌کرد. او در سال ۲۰۲۰ مدرک دکترای خود را در زیست‌شناسی محاسباتی گرفت، جایی که او ابزارهایی را برای انتساب RNA-seq تک‌سلولی با استفاده از پیشین‌ها و تجزیه‌وتحلیل یکپارچه داده‌های چندگانه Omics با استفاده از یادگیری عمیق منتشر کرد.

[11] https://github.com/jonathanronen

فصل ۱

مقدمه‌ای بر ژنومیک

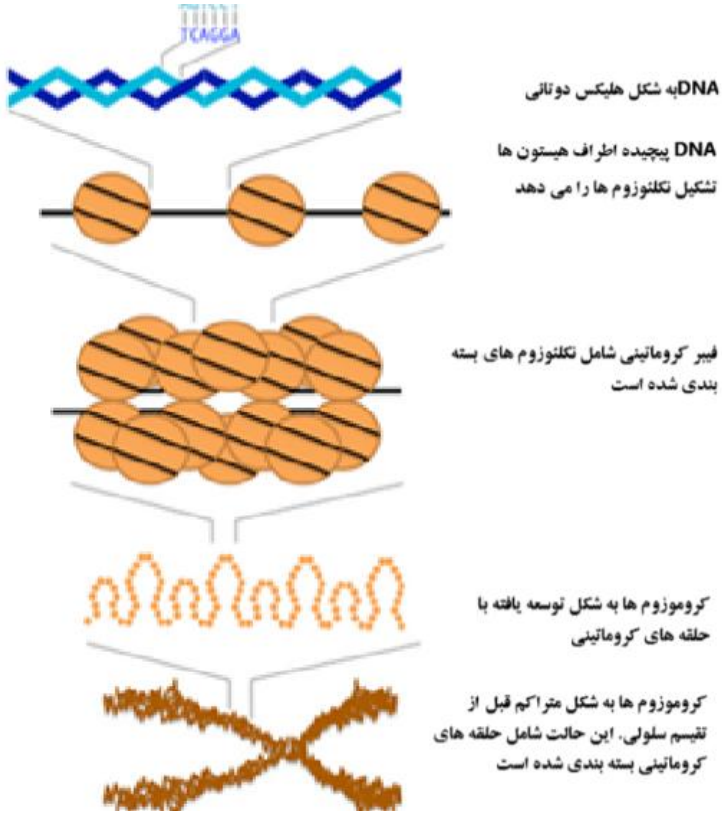
هدف این فصل ارائه برخی از اصول موردنیاز برای درک زیست‌شناسی ژنوم در اختیار خواننده است. به‌هیچ‌وجه، این یک مرور کلی از موضوع نیست، بلکه فقط خلاصه‌ای است که به خواننده غیر زیست‌شناس کمک می‌کند تا مفاهیم بیولوژیکی تکرارشونده در ژنومیک محاسباتی را درک کند. خوانندگانی که در زیست‌شناسی ژنوم و سنجش‌های کمی ژنومی مدرن به‌خوبی مسلط هستند، باید از این فصل بگذرند یا آن را مرور کنند.

۱-۱ ژن‌ها، DNA و اصل مرکزی

یک مفهوم اصلی که بارهاوبارها مطرح می‌شود "ژن" است. قبل از اینکه بتوانیم آن را توضیح دهیم، باید چند مفهوم دیگر را معرفی کنیم که برای درک مفهوم ژن مهم هستند. بدن انسان از میلیاردها سلول تشکیل شده است. این سلول‌ها در کارهای مختلفی تخصص دارند. به‌عنوان‌مثال، در کبد سلول‌هایی وجود دارد که به تولید آنزیم‌هایی برای شکستن سموم کمک می‌کنند. در قلب، سلول‌های ماهیچه‌ای تخصصی وجود دارد که باعث تپش قلب می‌شود. بااین‌حال، همه این انواع مختلف سلول‌ها از یک جنین تک‌سلولی می‌آیند. تمام دستورالعمل‌های ساخت انواع مختلف سلول در آن سلول وجود دارد و با هر تقسیم آن سلول، این دستورالعمل‌ها به سلول‌های جدید منتقل می‌شوند. این دستورالعمل‌ها می‌توان در یک‌رشته - یک مولکول DNA، یک پلیمر ساخته شده از واحدهای تکرارشونده به نام نوکلئوتید رمزگذاری کرد. چهار نوکلئوتید موجود در مولکول‌های DNA، آدنین، گوانین، سیتوزین و تیمین (با چهار حرف رمزگذاری شده‌اند: A، C، G و T) در یک توالی خاص، اطلاعات را برای زندگی ذخیره می‌کنند. DNA به شکل دو مارپیچ سازماندهی شده است که در آن دو پلیمر مکمل با یکدیگر درهم می‌پیچند و به شکل مارپیچ آشنا می‌پیچند.

۱-۱-۱ ژنوم چیست؟

توالی کامل DNA یک موجود زنده که حاوی تمام اطلاعات ارثی است، ژنوم نامیده می‌شود. ژنوم شامل تمام اطلاعات برای ساخت و نگهداری ارگانیسم است. ژنوم‌ها در اندازه‌ها و ساختارهای مختلف هستند. ژنوم ما تنها یک بخش خالص DNA نیست. در سلول‌های یوکاریوتی، DNA به‌دور پروتئین‌ها (هیستون‌ها) پیچیده می‌شود و ساختارهای مرتبه بالاتری مانند نوکلئوزوم‌ها را تشکیل می‌دهند که کروماتین‌ها و کروموزوم‌ها را می‌سازند (شکل ۱-۱ را ببینید).



شکل ۱-۱: ساختار کروموزوم در حیوانات.

بسته به ارگانیسم ممکن است چندین کروموزوم وجود داشته باشد. بااین‌حال، در برخی از گونه‌ها (مانند اکثر پروکاریوت‌ها) DNA به شکل دایره‌ای ذخیره می‌شود. اندازه ژنوم بین گونه‌ها نیز متفاوت است. ژنوم انسان دارای ۴۶ کروموزوم و بیش از ۳ میلیارد جفت‌پایه است، درحالی‌که ژنوم گندم دارای ۴۲ کروموزوم و ۱۷ میلیارد جفت‌پایه است. هم اندازه ژنوم و هم تعداد کروموزوم‌ها بین موجودات مختلف متغیر است. توالی ژنوم موجودات با استفاده از فناوری توالی‌یابی به دست می‌آید. با این فناوری، قطعاتی از توالی DNA از ژنوم به دست می‌آید که خوانش نامیده می‌شود. تکه‌های بزرگ‌تری از توالی ژنوم بعداً با اتصال قطعات اولیه به قطعات بزرگ‌تر با استفاده از همپوشانی خوانش‌ها به دست می‌آید. جدیدترین فناوری‌های توالی‌یابی، توالی‌یابی ژنوم را ارزان‌تر و سریع‌تر کرده است. این فناوری‌ها تولید خوانش بیشتر، خوانش طولانی‌تر و خوانش دقیق‌تر می‌کنند.

هزینه تخمینی اولین ژنوم انسان ۳۰۰ میلیون دلار در سالهای ۱۹۹۹-۲۰۰۰ بود. امروزه با ۱۵۰۰ دلار می‌توان یک ژنوم انسانی باکیفیت بالا به دست آورد. ازآنجایی‌که هزینه‌ها در حال کاهش است، محققان و پزشکان می‌توانند داده‌های بیشتری تولید کنند. این امر هزینه‌های ذخیره‌سازی داده‌ها را افزایش می‌دهد و همچنین تقاضا برای افراد واجد شرایط برای تجزیه‌وتحلیل داده‌های ژنومی را افزایش می‌دهد. این یکی از انگیزه‌های نگارش این کتاب بود.

۱-۱-۲ ژن چیست؟

در ژنوم، مناطق خاصی حاوی اطلاعات دقیقی وجود دارد که محصولات فیزیکی اطلاعات ژنتیکی را رمزگذاری می‌کند. منطقه‌ای در ژنوم با این اطلاعات به طور سنتی "ژن" نامیده می‌شود. بااین‌حال، تعریف دقیق ژن هنوز درحال‌توسعه است. طبق کتاب‌های درسی کلاسیک زیست‌شناسی مولکولی، یک ژن بخشی از یک توالی DNA است که مرتبط به یک پروتئین منفرد یا یک مولکول RNA ساختاری و کاتالیزوری است (Alberts et al., 2002). یک تعریف مدرن این است: "یک منطقه (مناطق) که شامل تمام عناصر دنباله‌ای لازم برای رمزگذاری رونوشت عملکردی است (Eilbeck