بسم الله الرحمن الرحیم

انتشارات
۷۹۰

تجزیه‌وتحلیل داده‌های RNA-seq

همراه با لوح فشردهٔ آموزش گام‌به‌گام

دکتر علیرضا سیفی
عضو هیئت علمی دانشگاه فردوسی مشهد
مهندس محمدرضا رضائی

سرشناسه: سیفی، علیرضا، ۱۳۵۶-
عنوان و نام پدیدآور: تجزیه و تحلیل داده‌های RNA-seq همراه با لوح فشرده آموزش گام‌به‌گام/ علیرضا سیفی، محمدرضا رضایی؛ ویراستار علمی امین میرشمسی‌کاخکی؛ ویراستار ادبی هانیه اسدپور فعال مشهد.
مشخصات نشر: مشهد: دانشگاه فردوسی مشهد، ۱۴۰۰.
مشخصات ظاهری: ۱۲۸ ص.
فروست: انتشارات دانشگاه فردوسی مشهد؛ ۷۹۰.
شابک: ISBN: 978-964-386-483-5
وضعیت فهرست‌نویسی: فیپا.
موضوع: آر. ان. ا. RNA
موضوع: نوکلئوتیدها -- توالی Nucleotide sequence
شناسه افزوده: رضایی، محمدرضا، ۱۳۷۳-
شناسه افزوده: میرشمسی کاخکی، امین، ۱۳۶۵-، ویراستار
شناسه افزوده: دانشگاه فردوسی مشهد، انتشارات.
رده‌بندی کنگره: QP۶۲۳
رده‌بندی دیویی: ۵۷۲/۸۸
شماره کتابشناسی ملی: ۷۶۱۴۲۶۹

تجزیه‌وتحلیل داده‌های RNA-seq

همراه با لوح فشردهٔ آموزش گام‌به‌گام

پدیدآورندگان: دکتر علیرضا سیفی؛ مهندس محمدرضا رضائی
ویراستار علمی: دکتر امین میرشمسی کاخکی
ویراستار ادبی: هانیه اسدپور فعال مشهد
مشخصات: وزیری، ۲۰۰ نسخه، چاپ اوّل، پاییز ۱۴۰۰
چاپ و صحافی: چاپخانهٔ دقت
بها : ۳۲۰٫۰۰۰ ریال

انتشارات
۷۹۰

مراکز پخش:

فروشگاه و نمایشگاه کتاب پردیس: مشهد، میدان آزادی، دانشگاه فردوسی مشهد، جنب سلف یاس تلفن: ۳۸۸۰۲۶۶۶- ۳۸۸۳۳۷۲۷ (۰۵۱)
مؤسّسهٔ کتابیران: تهران، خیابان کارگر جنوبی، خیابان لبافی‌نژاد، بین خیابان فروردین و اردیبهشت، شمارهٔ ۲۳۸، تلفن: ۶۶۴۹۴۴۰۹-۶۶۴۸۴۷۱۵ (۰۲۱)
مؤسّسهٔ دانشیران: تهران، خیابان انقلاب، خیابان منیری جاوید (اردیبهشت) نبش خیابان نظری، شمارهٔ ۱۴۲ تلفکس: ۶۶۴۰۰۲۲۰ – ۶۶۴۰۰۱۴۴ (۰۲۱)

http://press.um.ac.ir Email: press@um.ac.ir

فهرست مطالب

پیشگفتار ۷

فصل ۱. تاریخچهٔ پیشرفت‌ها در ژنتیک، بیولوژی مولکولی و بیوانفورماتیک ۹

نسل دوم روش‌های توالی‌یابی ۱۲

نسل سوم روش‌های توالی‌یابی ۱۵

ابزارهای آنالیز داده‌های NGS ۱۶

منابع ۱۹

فصل ۲. آشنایی با اصول کامپیوتر برای پژوهشگران علوم زیستی ۲۱

انتخاب سیستم کامپیوتری مناسب برای آنالیز داده‌های NGS ۲۲

حافظه‌های ذخیره‌سازی ثانویه ۲۴

ریزپردازندهٔ مرکزی یا CPU ۲۵

حافظهٔ اصلی یا RAM ۲۶

منابع ۲۶

فصل ۳. سیستم‌عامل لینوکس ۲۷

تاریخچهٔ نرم‌افزارهای با دسترسی آزاد و پیدایش لینوکس ۲۷

نصب سیستم‌عامل Ubuntu ۲۸

دستورات در Ubuntu ۲۹

کلیدهای ترکیبی در خط فرمان لینوکس ۳۳
منابع ۳۴

فصل ۴. محیط نرم‌افزاری و زبان برنامه‌نویسی R ۳۵
نصب R ۳۷
توابع R ۳۹
ساده‌سازی با ایجاد متغیر ۴۰
پوشهٔ کاری و تغییر آن ۴۰
داده‌ها در R ۴۱
ساختار داده‌ها در R ۴۲
روش‌های وارد کردن داده‌ها به R ۴۵
عملیات ریاضی روی وکتورها ۴۷
مقایسات منطقی ۴۷
عملگرهای منطقی ۴۸
ترسیم نمودارهای آماری در R ۴۹
نمودار هیستوگرام ۵۰
بستهٔ نرم‌افزاری ggplot2 ۵۱
برخی دیگر از توابع مهم R ۵۵
نوشتن برنامه در R ۵۶
منابع ۵۸

فصل ۵. RNA-seq: انقلابی در مطالعهٔ ترنسکریپتوم ۵۹
روش‌های ارزیابی بیان ژن ۵۹
طراحی آزمایشات RNA-seq ۶۱
روش‌های نوین RNA-seq ۶۵
استفاده از داده‌های RNA-seq موجود در NCBI ۶۶

منابع ۶۸

فصل ۶. ارزیابی کیفیت و پردازش داده‌های ایلومینا ۷۰
فایل fastq ۷۰
کنترل کیفیت خوانش‌ها با استفاده از ابزار FASTQC ۷۴
نصب نرم‌افزار FASTQC ۷۴
پردازش توالی‌ها با استفاده از نرم‌افزار Trimmomatic ۷۹
منابع ۸۳

فصل ۷. هم‌ردیف‌کردن خوانش‌ها با توالی مرجع ۸۶
مسیرهای مختلف آنالیز داده‌های RNA-seq ۸۶
روش‌های هم‌ردیف کردن خوانش‌ها با توالی مرجع ۸۸
هم‌ردیفی روی ژنوم مرجع با استفاده از HISAT2 ..... ۸۹
هم‌ردیف کردن خوانش‌ها با استفاده از BWA ۹۰
جدول تعداد خوانش‌های هم‌ردیف‌شده ۹۲
فیلتر کردن جدول خوانش‌های هم‌ردیف‌شده ۹۲
منابع ۹۳

فصل ۸. آنالیز بیان ژن‌ها ۹۴
بررسی اوّلیهٔ خوانش‌های هم‌ردیف‌شده ۹۴
شناسایی ژن‌های با بیان متفاوت با استفاده از DESeq2 ..... ۹۶
ترسیم نقشهٔ حرارتی ۹۸
بررسی بیان ژن‌ها با استفاده از StringTies-Ballgown ۱۰۰
تجزیه‌وتحلیل‌های پس از شناسایی ژن‌های با بیان متفاوت ۱۰۲
تهیهٔ ترنسکریپتوم مرجع ۱۰۲
منابع ۱۰۴

فصل ۹. ملاحظات آماری در تجزیه‌وتحلیل آزمایشات RNA-seq ۱۰۵
طرح آزمایشی در مطالعات RNA-seq ۱۰۵
آزمایش بدون تکرار زیستی ۱۰۶
نرمال‌سازی تعداد خوانش‌های هم‌ردیف‌شده ۱۰۷
مدل سازی خوانش‌های RNA-seq ۱۰۹
یادآوری ۱۱۰
یادآوری ۱۱۱
منابع ۱۱۲

پیوست: نصب لینوکس مجازی روی سیستم‌عامل ویندوز ۱۱۳
نمایه ۱۳۴

پیشگفتار

فنّاوری‌های نوین توالی‌یابی DNA که با عنوان متداول NGS شناخته می‌شوند، تحول شگرفی در رشته‌های مختلف علوم زیستی ایجاد کرده و امکان مطالعهٔ سریع‌تر و جامع‌تر اساس ژنتیکی و مولکولی پدیده‌های زیستی را فراهم آورده‌اند. هم‌راستا با موفقیت‌های چشمگیر سخت‌افزاری در فنّاوری‌های توالی‌یابی DNA، پیشرفت‌های اساسی نیز در تولید ابزارهای تجزیه‌وتحلیل داده‌های حاصل از این فنّاوری‌ها حاصل شده است.

چالش اصلی در تجزیه‌وتحلیل داده‌های NGS حجم بالا و پیچیدگی این داده‌هاست که بهره‌برداری مناسب از آن ها نیازمند استفاده از ابزارهای بیوانفورماتیکی کارا و دقیق است. این ابزارهای بیوانفورماتیکی معمولاً توسط متخصّصان علوم ریاضی، کامپیوتر و بیوانفورماتیک ساخته می‌شوند و در غالب موارد بهره‌برداری از آن‌ها برای متخصّصان علوم زیستی که اطلاعات کافی از علوم کامپیوتری ندارند، چالش‌برانگیز است. معمولاً در مراکز تحقیقاتی پویا در دنیا یک یا چند متخصّص بیوانفورماتیک عهده‌دار تجزیه‌وتحلیل داده‌های NGS تولیدشده توسط محقّقان زیست‌شناسی هستند و بنابراین تمام محقّقان نیازمند فراگیری این روش‌های تجزیه‌وتحلیل نیستند. لیکن در ایران به‌دلیل اینکه هنوز توسعه‌یافتگی مطلوبی در مراکز تحقیقاتی رخ نداده است و کماکان تشکیل تیم‌های تحقیقاتی با تخصص‌های گوناگون میسر نیست، محقّقان علوم زیستی نیاز دارند که اشراف نسبی بر روش های تجزیه‌وتحلیل NGS داشته باشند.

مطالعهٔ ترنسکریپتوم (کل محتوای RNA سلول) با استفاده از روش‌های توالی‌یابی RNA (که با عنوان RNA-seq شناخته می‌شود) یکی از قدرتمندترین روش‌های مطالعهٔ پدیده‌های زیستی برای شناسایی مکانیسم‌های مولکولی و ژنتیکی کنترل‌کنندهٔ این پدیده‌هاست. ابزارهای بیوانفورماتیکی بسیار متنوعی برای تجزیه‌وتحلیل داده‌های RNA-seq ارائه شده است و ابزارهای جدید نیز با سرعت درحال توسعه و ارائه است. هدف از گردآوری کتاب حاضر فراهم‌آوردن مجموعه‌ای است برای محقّقان علوم زیستی که الزاماً اطلاعات گسترده‌ای از بیوانفوماتیک ندارند، ولی نیازمند تجزیه‌وتحلیل داده‌های RNA-seq هستند.

در این کتاب مباحث نظری موردنیاز درمورد RNA-seq و اصول پایه و مقدماتی کامپیوتر توضیح داده می‌شود. سپس روش‌های استاندارد تجزیه‌وتحلیل داده‌های RNA-seq که درحال‌حاضر بیشترین کاربرد را در بین پژوهشگران دارند، معرفی می‌شوند. با استفاده از داده‌های واقعی کل مسیرهای تجزیه‌وتحلیل به‌صورت گام‌به‌گام توضیح داده می‌شوند، به‌نحوی‌که مخاطب با دراختیارداشتن این کتاب بتواند کل مسیر تجزیه‌وتحلیل RNA-seq را روی کامپیوتر شخصی خود تمرین کند. کلّیهٔ نرم‌افزارها، داده‌های مورداستفاده

و خروجی‌های موردانتظار از اجرای هرکدام از مراحل تجزیه‌وتحلیل در قالب یک لوح فشرده همراه کتاب در اختیار خوانندگان قرار می‌گیرد. تلاش شده است که مطالب کتاب و لوح فشرده همراه کتاب به‌نحوی تنظیم شود که علاقه‌مندان به فراگیری روش‌های تجزیه‌وتحلیل RNA-seq با صرف حداقل زمان، اصول این روش‌ها را دریابند و بتوانند از آن‌ها در پروژه‌های پژوهشی خود استفاده کنند.

مزیت عمدهٔ این کتاب این است که ابزارهایی که در آن معرفی می‌شوند ابزارهای استاندارد، رایگان و با متن‌باز هستند. استفاده از این ابزارها به پرداخت هزینه و یا عدول از اصول اخلاقی مربوط به حقوق مالکیت معنوی نیاز ندارد. چنانچه تجزیه‌وتحلیل با این ابزارها فراگرفته شود، استفاده از ابزارهای جدیدتری که به‌سرعت درحال توسعه و توزیع در جامعهٔ علمی هستند، بسیار ساده‌تر خواهد بود و پژوهشگر همواره به جدیدترین ابزارهای تجزیه‌وتحلیل RNA-seq دسترسی خواهد داشت.

هرچند تلاش زیادی شده است که اشکالات نگارشی و علمی در متن وجود نداشته باشد، بااین‌حال سپاسگزار خواهم بود چنانچه خوانندگان محترم اشکالات احتمالی و یا پیشنهادهای سازندهٔ خود را برای ارتقای کیفی کتاب به‌نحو مقتضی به این‌جانب منعکس کنند.

علیرضا سیفی
زمستان ۱۳۹۹