

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



بیوانفورماتیک

کتاب راهنمای عملی توالی‌یابی نسل جدید و کاربردهای آن

برای دریافت تصاویر رنگی و فایل ضمیمه (داده‌های همراه)، به پروفایل کتاب در تارنمای انتشارات دانشگاه فردوسی مشهد به نشانی زیر مراجعه فرمایید:

press.um.ac.ir

لوید لو - مارتی تامی

ترجمه:

دکتر امین میرشمسی کاخکی

استادیار دانشگاه فردوسی مشهد

دکتر احمد جلیلیان

عنوان و نام پدیدآور:	بیوانفورماتیک؛ کتاب راهنمای عملی توالی‌یابی نسل جدید و کاربردهای آن/ لوید و مارتی تامی؛ ترجمه امین میرشمسی کاخکی، احمد جلیلیان؛ ویراستار علمی علیرضا سیفی؛ ویراستار ادبی هانیه اسدپور فعال مشهد.
مشخصات نشر:	مشهد: دانشگاه فردوسی مشهد، انتشارات، ۱۴۰۲.
مشخصات ظاهری:	۲۰۴ ص.: جدول، نمودار
فروست:	انتشارات دانشگاه فردوسی مشهد؛ ۸۹۸.
شابک:	
وضعیت فهرست‌نویسی:	فیپا.
یادداشت:	عنوان اصلی: Bioinformatics : a practical handbook of next generation sequencing and its applications, c2017
یادداشت:	کتاب حاضر قبلاً با عنوان "راهنمای عملی آنالیز داده‌های توالی‌یابی نسل جدید (NGS) و کاربردهای آن" با ترجمه محمد محمودی گماری... و دیگران در سال ۱۴۰۱ توسط انتشارات رسانه تخصصی ترجمه و منتشر شده است.
یادداشت:	کتابنامه. نمایه.
عنوان دیگر:	راهنمای عملی آنالیز داده‌های توالی‌یابی نسل جدید (NGS) و کاربردهای آن.
موضوع:	زیست‌انفورماتیک نوکلئوتیدها — توالی زیست‌شناسی مولکولی ژن نگاشت
شناسه افزوده:	لو، لوید، ویراستار
شناسه افزوده:	تامی، مارتی تی، ویراستار
شناسه افزوده:	میرشمسی کاخکی، امین، ۱۳۶۵ - مترجم
شناسه افزوده:	جلیلیان، احمد، ۱۳۶۷ - مترجم
شناسه افزوده:	سیفی، علیرضا، ۱۳۵۶ - ویراستار
شناسه افزوده:	دانشگاه فردوسی مشهد، انتشارات.
رده‌بندی کنگره:	QH۲۲۴/۲
رده‌بندی دیویی:	۵۷۰/۲۸۵
شماره کتابشناسی ملی:	۹۴۴۵۷۱۰

بیوانفورماتیک؛ کتاب راهنمای عملی توالی‌یابی نسل جدید و کاربردهای آن

پدیدآورندگان: لوید لو - مارتی تامی
ترجمه: دکتر امین میرشمسی کاخکی، دکتر احمد جلیلیان
ویراستار علمی: دکتر علیرضا سیفی
ویراستار ادبی: هانیه اسدپور فعال مشهد
مشخصات: وزیری، ۲۰۰ نسخه، چاپ اول، زمستان ۱۴۰۲
چاپ و صحافی: چاپخانه دقت
بها: ۱/۵۰۰/۰۰۰ ریال
حق چاپ برای انتشارات دانشگاه فردوسی مشهد محفوظ است.

مراکز پخش:

فروشگاه و نمایشگاه کتاب پردیس: مشهد، میدان آزادی، دانشگاه فردوسی مشهد، جنب سلف یاس
تلفن: ۳۸۸۰۲۶۶۶ - ۳۸۸۳۳۷۲۷ (۰۵۱)
مؤسسه کتابیران: تهران، میدان انقلاب، خیابان کارگر جنوبی، بین روانمهر و وحید نظری، بن‌بست
گشتاسب، پلاک ۸ تلفن: ۶۶۴۸۴۷۱۵ (۰۲۱)
مؤسسه دانشیران: تهران، خیابان انقلاب، خیابان منیری جاوید (اردیبهشت) نبش خیابان نظری، شماره ۱۴۲
تلفکس: ۶۶۴۰۰۲۲۰ - ۶۶۴۰۰۱۴۴ (۰۲۱)

<http://press.um.ac.ir>

Email: press@um.ac.ir



انتشارات
۸۹۸

فهرست مطالب

پیشگفتار مترجمان	۹
پیشگفتار مؤلف	۱۰
پیشگفتار	۱۲
مقدمه	۱۴
فصل ۱. آشنایی با فناوری‌های توالی‌یابی نسل جدید	۱۵
تاریخچه مختصری از توالی‌یابی DNA	۱۵
فناوری‌های توالی‌یابی نسل جدید	۱۷
پلت فرم 454	۱۹
پلت فرم ABI SOLiD	۲۱
پلت فرم ایلومینا	۲۳
پلت فرم Ion Torrent	۲۶
پلت فرم Pacific Biosciences	۲۷
پلت فرم Oxford Nanopore Technologies	۲۸
چالش‌های انفورماتیکی	۲۹
منابع	۳۰
فصل ۲. مبانی لینوکس	۳۳
مقدمه	۳۳
فهرست کردن محتوای یک دایرکتوری	۳۳
ایجاد دایرکتوری	۳۶
چاپ مسیر دایرکتوری کار	۳۶
تغییر دایرکتوری	۳۷
دانلود داده‌ها	۳۷
فشرده‌سازی فایل	۳۸
نمایش محتوای یک فایل	۳۹
شمارش تعداد خطوط	۴۰
جست‌وجوی یک الگو	۴۰

۴۱	ترکیب چندین دستور مختلف با هم
۴۱	تبدیل یک فایل FASTQ به فرمت Tabular
۴۳	تطبیق الگو با استفاده از Awk
۴۴	مرتب‌سازی و استخراج توالی‌های منحصربه‌فرد
۴۵	تبدیل خوانش‌ها به توالی‌های فرمت FASTA
۴۷	نوشتن یک اسکریپت Shell برای جدا کردن توالی‌ها به فایل‌های جداگانه
۴۹	تغییر مجوزهای فایل
۵۰	اجرای اسکریپت Bash
۵۱	خلاصه

۵۳	فصل ۳. بررسی کیفیت توالی
۵۳	مقدمه
۵۴	FastQC
۵۴	مرحله نصب در محیط لینوکس
۵۶	دانلود مجموعه داده‌ها
۶۴	ابزارهای پردازش FASTQ و Fastx-Toolkit
۶۴	مرحله نصب در محیط لینوکس
۶۸	نتیجه‌گیری
۶۸	منابع

۶۹	فصل ۴. هم‌ردیفی خوانش‌های توالی‌یابی شده
۶۹	مقدمه
۷۰	عملی
۷۰	هم‌ردیفی خوانش‌های کوتاه
۷۱	فرایند هم‌ردیفی
۷۱	ایجاد ایندکس توالی مرجع
۷۱	ایجاد ایندکس مرجع با BWA
۷۱	هم‌ردیفی خوانش‌ها با مرجع
۷۱	هم‌ردیفی خوانش‌ها با BWA
۷۱	مشاهده فایل SAM
۷۵	تبدیل فرمت SAM به BAM
۷۵	مرتب کردن هم‌ردیف‌های BAM

۷۶	ابزار پیشنهادی جایگزین: novoAlign & novoSort
۷۶	ایجاد ایندکس novoAlign
۷۶	هم‌ردیفی خوانش‌ها با novoAlign
۷۶	تبدیل SAM به BAM
۷۶	مرتب کردن و حذف خوانش‌های مضاعف‌شده با استفاده از novoSort
۷۷	مشاهده هم‌ردیفی BAM با IGV
۷۷	اجرای IGV
۷۷	منابع

۷۹	فصل ۵. ایجاد یک گردش کار تحقیقاتی
۷۹	مقدمه
۸۰	داده‌های خام
۸۱	اسکرپت‌های Shell
۸۵	گلکسی
۸۵	۱- صفحه وب گلکسی را باز کنید
۸۵	۲- ایجاد حساب گلکسی
۸۷	۳- گرفتن داده‌ها
۸۹	۴- اجرای FastQC قبل از پیرایش
۹۱	۵- اجرای Trimmomatic
۹۳	۶- اجرای FastQC بعد از پیرایش
۹۴	۷- ایجاد گردش کار از تاریخچه
۱۰۰	۸- اجرای گردش کار
۱۰۱	نتیجه‌گیری
۱۰۲	منابع

۱۰۳	فصل ۶. سرهم‌بندی کردن De novo یک ژنوم
۱۰۴	مقدمه
۱۰۶	مراحل کلی
۱۰۷	۱- دانلود توالی‌ها
۱۰۸	۲- فیلتر کردن خوانش‌های بد
۱۰۹	۳- ا. سرهم‌بندی خوانش‌های کوتاه جفت انتهایی
۱۱۰	۳- ب. سرهم‌بندی هیبرید با خوانش‌های PacBio

۱۱۱	۳- ج. سرهم‌بندی کردن خوانش تک انتهای بلند
۱۱۲	۴- بررسی کیفیت ژنوم
۱۱۲	آمار
۱۱۴	تکاملی
۱۱۶	بحث و نتیجه‌گیری
۱۱۷	منابع

۱۱۹	فصل ۷. توالی‌یابی اگزوم
۱۱۹	مقدمه
۱۲۰	گردش کار کلی WES
۱۲۱	اطلاعات پیش‌زمینه درمورد بخش عملی
۱۲۲	نرم‌افزار
۱۲۲	مجموعه داده‌ها
۱۲۳	دانلود مجموعه داده‌ها
۱۲۳	ایجاد پوشه جدید
۱۲۳	مکان‌یابی داده‌های خام روی ژنوم مرجع
۱۲۴	فراخوان واریانت‌ها
۱۲۶	پیش‌بینی اثرات SNV ها و Indels
۱۲۷	نمایش گرافیکی
۱۲۸	نتیجه‌گیری
۱۲۹	منابع

۱۳۱	فصل ۸. ترانسکریپتومیکس
۱۳۱	مقدمه
۱۳۲	اهداف
۱۳۳	مجموعه داده‌ها و نرم‌افزارها
۱۳۳	پیش‌پردازش و کنترل کیفیت (QC) خوانش‌ها
۱۳۳	آماده‌سازی فایل‌ها
۱۳۴	اجرای ساده QC خوانش‌ها
۱۳۴	پیرایش آداپتور
۱۳۵	STAR: هم‌ردیفی خوانش‌ها
۱۳۵	آماده‌سازی فایل‌ها

۱۳۵.....	ایندکس ژنوم مرجع
۱۳۶.....	هم‌ردیفی خوانش‌ها
۱۳۶.....	بیان افتراقی
۱۳۷.....	بیان ژن مبتنی بر شمارش با استفاده از HTSeq-count و edgeR
۱۳۷.....	HTSeq-count
۱۳۹.....	edgeR: بیان ژن و بیان افتراقی
۱۳۹.....	آماده‌سازی فایل‌ها: ادغام جداول HTSeq-count
۱۳۹.....	بیان ژن افتراقی با استفاده از ابزار edgeR
۱۴۳.....	بیان ژن مبتنی بر FPKM با استفاده از بسته نرم‌افزاری Cufflinks
۱۴۳.....	آنالیز بیان افتراقی با استفاده از Cufflinks
۱۴۳.....	ادغام Cufflinks Assembly با استفاده از Cuffmerge
۱۴۴.....	کمی کردن بیان با استفاده از Cuffquant
۱۴۴.....	آنالیز بیان افتراقی با استفاده از Cuffdiff
۱۴۶.....	نمایش نتایج با بسته نرم‌افزاری cummeRbund
۱۴۹.....	منابع

۱۵۱.....	فصل ۹. متاژنومیکس
۱۵۱.....	مقدمه
۱۵۲.....	آشنایی با گردش کار سرور MG-RAST
۱۵۲.....	ثبت نام در MG-RAST
۱۵۳.....	بارگذاری مجموعه داده‌ها
۱۵۴.....	پایش وضعیت کاری
۱۵۴.....	مشاهده و آنالیز نتایج
۱۵۴.....	نصب QIIME
۱۵۸.....	متاژنومیکس 16S rRNA
۱۵۸.....	شروع کار
۱۵۹.....	بارگذاری و ارسال
۱۶۲.....	نتایج
۱۶۵.....	توالی‌یابی متاژنومی شاتگان
۱۶۵.....	شروع کار
۱۶۷.....	بارگذاری و ارسال
۱۶۷.....	نتایج

۱۶۹.....	ایجاد جدول تاکسونومی BIOM
۱۷۱.....	ایجاد جدول BIOM عملکردی
۱۷۲.....	نتیجه گیری
۱۷۳.....	منابع
۱۷۵.....	فصل ۱۰. کاربرد داده‌های توالی‌یابی نسل جدید
۱۷۶.....	مقدمه
۱۷۸.....	نقشه پیوستگی کلاسیک
۱۷۹.....	OneMap (2.0-0)
۱۸۰.....	نصب و راه‌اندازی
۱۸۰.....	فرمت ورودی
۱۸۱.....	آنالیز نقشه‌یابی لینکاژ
۱۸۶.....	از نقشه پیوستگی ژنی تا نقشه فیزیکی
۱۸۹.....	مطالعات ارتباطی جامع ژنومی (GWAS)
۱۸۹.....	PLINK 1.90 Beta
۱۸۹.....	نصب و راه‌اندازی
۱۹۰.....	فایل‌های ورودی و فرمت
۱۹۰.....	دانلود مجموعه داده‌ها
۱۹۱.....	آنالیز ارتباطی
۲۰۲.....	نتیجه گیری
۲۰۲.....	منابع
۲۰۳.....	نمایه

پیشگفتار مترجمان

تحولات بی‌پایان تکنولوژی به پیشرفت‌های قابل توجه در تعیین توالی ژنوم منجر شده است. امروزه توالی‌یابی نسل جدید (NGS) به‌طور چشمگیری توسعه یافته و از نظر اقتصادی مقرون به‌صرفه شده است. از طرفی، حجم بالای داده‌های توالی‌یابی که روزانه تولید می‌شود، ضرورت توسعه روش‌های آنالیز، تحلیل و ارتقای مهارت‌ها در این خصوص را جدی‌تر می‌کند. مهارت‌های مورد نیاز برای تجزیه و تحلیل محاسباتی درمورد طیف گسترده‌ای از برنامه‌های تحقیقاتی شامل شناسایی مکانیسم‌های مولکولی پاسخ به تنش‌های زیستی و غیرزیستی، تولید گیاهان زراعی با عملکرد بالاتر، شناسایی علل بیماری‌ها، طراحی واکسن، آنتی‌بیوتیک‌های جدید، تولید دارو و غیره بسیار مورد توجه است. این کتاب راهنمای گام به گام حل موضوعات پیچیده‌ای را ارائه می‌دهد که در نهایت فرایند انجام تجزیه و تحلیل، از داده‌های خام توالی تا پاسخ به سؤالات مهم بیولوژیکی را برای خوانندگان آسان می‌کند. فصل‌های کتاب به‌صورت دستورالعمل‌های مستقل طراحی و نگارش شده‌اند و این امر برای خوانندگانی که به یادگیری برخی موضوعات خاص تمایل دارند، کاملاً مناسب است. این کتاب شامل مباحث کلی درمورد چگونگی مدیریت داده‌های NGS از بررسی کیفیت توالی تا سرهم‌بندی خوانش‌ها برای یافتن چندشکلی‌های تک‌نوکلئوتیدی (SNP) است. سایر موضوعات پیشرفته مانند سرهم‌بندی ژنوم، تعیین توالی اگزوم، آنالیز ترانسکریپتوم، متاژنومیکس و همچنین جنبه‌های کاربردی در نقشه‌یابی ژنتیکی و مطالعه ارتباطی جامع ژنومی (GWAS) نیز در این کتاب پوشش داده شده‌اند. قابل ذکر است که مترجمان علاوه بر کتاب حاضر، کتاب‌های مشابه دیگر را مطالعه و بررسی کرده و به جرئت ادعا می‌کنند که هیچ کدام از آن‌ها به اندازه اثر حاضر تا این حد روش‌های تجزیه و تحلیل داده‌های حاصل از توالی‌یابی را به زبان ساده بیان نکرده‌اند. درواقع برای کسانی که می‌خواهند به تازگی با دنیای تجزیه و تحلیل داده‌های NGS آشنا شوند، مطالعه این کتاب توصیه می‌شود. این کتاب علاوه بر داشتن متنی ساده و روان، موضوعات تخصصی و روش‌های آنالیز را به‌صورت کاملاً کاربردی و ساده توضیح داده، به‌طوری که در انجام هر آنالیز منطق و اساس کار تا حد امکان توضیح داده شده است. به‌عنوان متخصص بیوتکنولوژی و مدرّس بیوانفورماتیک، این کتاب را یک همراه ضروری برای دانشجویان کارشناسی، دانشجویان تحصیلات تکمیلی رشته‌های مختلف زیستی، محققان و هر کسی که علاقه‌مند به یادگیری مؤثر در زمینه بیوانفورماتیک است، معرفی و توصیه می‌کنیم. مترجمان سعی کرده‌اند در ترجمه کتاب امانت‌دار باشند، اما اذعان می‌دارند که ترجمه این کتاب خالی از ایراد نیست. به‌همین خاطر از اساتید محترم، دانشجویان عزیز و سایر خوانندگان گرامی صادقانه خواستاریم که از این نقایص چشم‌پوشی نکنند و با دیدی نقّادانه کاستی‌های احتمالی را یادآوری کنند تا به‌عنوان خادمان نظام آموزش دانشگاه، آن‌ها را برطرف کنیم. همچنین در پایان از خانم دکتر سمیه حیدری که در ترجمه این کتاب ما را یاری کرده‌اند، سپاسگزاری می‌کنیم.

امید آن‌که این تلاش هرچند اندک مورد توجه و استفاده استادان، دانشجویان و پویندگان راه علم قرار گیرد.

مترجمان

تابستان ۱۴۰۲

پیشگفتار مؤلف

چندین سال پیش، هنگامی که توالی‌یابی اولین نسخه ژنوم انسان در حال تکمیل بود، مصمم شدم تلاش خود را بر روی مطالعه ژنوم پاتوژن‌ها متمرکز کنم. با داشتن تجربه‌ای در برنامه‌نویسی، یکی از اولین مواردی که ذهن من را مشغول کرد، مشکلی بود که در ابتدا با آن برخورد کردم و عملاً ارتباط چندانی با جنبه‌های جذاب زیست‌شناسی مطالعه ژنوم نداشت. از ابتدا مشخص بود که برای مطالعه تنوع ژنتیکی، اثرات آن بر فنوتیپ و بررسی پویایی تغییرات اپیدمیولوژیک مرتبط، لازم است حجم عظیم غیرقابل تصویری از داده‌ها را جمع‌آوری کنیم. چالش اصلی در این مسیر قابلیت ذخیره‌سازی داده‌ها یا قدرت پردازش موردنیاز نبود. قانون مور^۱ ما را به رشد سریع قدرت محاسبات عادت داده بود و من اطمینان داشتم که این چالش‌های فنی قابل برطرف شدن هستند.

سؤال مهم این بود که افرادی که این داده‌ها را تجزیه و تحلیل می‌کنند، آیا از دانش و منابع کافی برای رسیدگی به این حجم زیاد داده برخوردارند یا خیر؟ جهت اطمینان، بسیاری از شرکت‌های توسعه‌دهنده موتورهای جست‌وجو، نرم‌افزارهای رتبه‌بندی مبتنی بر وب و ارائه‌دهندگان خدماتی که ما اکنون آن را «داده‌های بزرگ» می‌نامیم نیز با مشکل مشابهی روبه‌رو بودند. با این حال، به نظر می‌رسید مشکل ژنومیکس به تنهایی توسط متخصصان علوم کامپیوتر قابل رفع نیست و از این جهت زیست‌شناسان مجبور بودند تا این حجم بسیار بالای داده‌ها را مدیریت کنند.

این مسائل حتی قبل از اینکه توالی‌یابی کل ژنوم با استفاده از فناوری‌های توالی‌یابی نسل جدید (NGS) توسط شرکت‌هایی مانند Solexa (ایلوئینا کنونی) معرفی شود، به ایجاد تحول شگرف در این زمینه منجر شده بود. امروزه، پروژه اپیدمیولوژی ژنومیک MalariaGEN^۲ که من روی آن مطالعه می‌کنم، شامل توالی‌یابی ژنوم انگل پلاسمودیوم^۳ از تقریباً ده هزار نمونه بالینی است که هریک چندین گیگابایت داده ایجاد کرده‌اند، داده‌هایی بسیار بیشتر از آنچه که چند سال پیش پیش‌بینی می‌کردم. بنابراین، هنوز چالش موجود رفع نشده و به‌طور جدی برای دانشمندان علوم زیستی و پزشکان تجزیه و تحلیل این حجم عظیم داده‌ها مشکل است. از طرف دیگر، بسیاری از پروژه‌ها بر همکاری با متخصصان انفورماتیک متکی هستند که اغلب تخصص محدودی در زمینه بیولوژی دارند.

1. Moore's Law

2. malariagen.net/projects/p-falciparum-community-project

3. Plasmodium

در سایه وجود این چالش‌ها، من از لوید لو^۱ و مارتی تامی^۲ به دلیل تلاش و سهم به‌سزایی که در رفع مشکلات ذکر شده داشتند، تقدیر می‌کنم. در واقع آنچه آن‌ها تهیه کرده‌اند، یک راهنمای بسیار کاربردی مهم به‌عنوان رفرنس و هم مرجع آموزشی است که به دلیل رویکرد منحصر به فرد آن توسط بسیاری از دانشمندان علوم زیستی مورد استقبال قرار خواهد گرفت. نکته مهم این است که محتوای این کتاب براساس سال‌ها تجربه تدریس بوده و با در نظر گرفتن مشکلات معمول افراد در مواجهه با داده‌های NGS تنظیم شده است. این کتاب شامل مجموعه تکنیک‌ها و شیوه‌های استفاده از برخی از محبوب‌ترین و مفیدترین ابزارهای مورد استفاده مانند BWA، Samtools و غیره است.

مطالب ارائه شده در این کتاب طیف وسیعی از جنبه‌های کاربردی را (به‌خصوص فصل آخر) پشتیبانی می‌کند، اما بدیهی است که هر خواننده باید با چالش‌های منحصر به فرد در زمینه مطالعه خود مواجه شود و نسبت به رفع آن اقدام کند. توصیف یک گردش کار تحلیلی منظم، جنبه‌های کیفی کار ژنومیکس که بسیار مهم است و اغلب در هنگام آنالیز دست کم گرفته می‌شود را برجسته می‌کند. آنالیز ارتباطی جامع ژنومی یا آنالیز فیلوژنی، بدون ارزیابی درست از داده‌ها، همواره به نتایج پیهوده یا کاذب منجر می‌شود. بنابراین تأکید و توصیه بر استانداردهای کیفی بالا در هنگام آنالیز برای دانشجویان و کسانی که تجزیه و تحلیل ژنومیکس را انجام می‌دهند، ضروری است. برای همه خوانندگان این کتاب آرزوی موفقیت در تلاش در این زمینه پیچیده علمی را دارم و امیدوارم پاداش‌های ارزشمندی نصیب آن‌ها شود.

دکتر اولیو میوتو^۳

واحد تحقیقات پزشکی گرمسیری ماهیدول-آکسفورد، بانکوک، تایلند

مؤسسه Wellcome Trust Sanger، هینکستون، انگلستان

مرکز ژنومیک و سلامت جهانی، دانشگاه آکسفورد، انگلستان

1. Lloyd Low

2. Martti Tammi

3. Olivo Miotto

پیشگفتار

انقلابی که روش‌های توالی‌یابی نسل جدید (NGS) در علم ژنتیک به ارمغان آورد را می‌توان با تحول علم نجوم به واسطه اختراع تلسکوپ مقایسه کرد. اکنون می‌توان پدیده‌های ژنتیکی را در مقیاس مولکولی، فرایندهای ژنتیکی را در سطوح ژنوم، ترانسکریپتوم و اپی‌ژنوم با استفاده از فناوری NGS مورد مطالعه قرار داد. هزینه پایین توالی‌یابی نسل جدید اجازه داده است تا توالی‌یابی ژنوم‌های انسانی به‌طور معمول انجام شود و توالی‌یابی ژنوم‌های بزرگ‌تر و ساخت آسان‌تر ژنوم‌های جدید امکان‌پذیر گردد. بنابراین، برای محققانی که در زمینه علوم زیستی فعالیت می‌کنند، ضروری است تا مفاهیم اساسی درخصوص مدیریت داده‌های NGS در سطوح مختلف را درک کنند. این کتاب به‌طور خلاصه و قابل‌فهم عموم مخاطبان را با روش‌های مختلف پردازش داده‌های NGS آشنا می‌کند.

برای کاربران مبتدی، سه فصل اول، مقدمه مختصری درخصوص فناوری NGS و نحوه رفع چالش‌های مرتبط با هم‌ردیفی داده‌های NGS با ژنوم مرجع را ارائه می‌دهد. مرحله هم‌ردیفی با استفاده از هم‌ردیف‌کننده محبوب BWA و هم‌ردیف‌کننده تجاری NovoAlign که به دقت بالا شناخته می‌شود، توضیح داده خواهد شد. این فصل کتاب توسط یکی از متخصصان شرکت نوواکرافت¹ نوشته شده است و خواننده می‌تواند با استفاده از ابزارهای تجاری یا گزینه‌های منبع باز، فرایند را برای دستیابی به درجه‌ای از دقت و سرعت مورد نیاز به‌دلوخواه تنظیم کند.

این کتاب مشکلات مربوط به هم‌ردیفی خوانش‌ها را مطرح می‌کند و به این سؤال معمول یعنی «در گام بعد چه کاری انجام دهم؟» پاسخ می‌دهد. کتاب با معرفی IGV به کاربران امکان تجسم هم‌ردیفی‌ها را می‌دهد و در مرحله بعد ابزار گلکسی را برای ایجاد یک گردش کار تحقیقاتی معرفی می‌کند. حتی اگر کاربر متخصص در علوم کامپیوتر نباشد، آشنایی با گلکسی به او این امکان را می‌دهد که پس از انجام آزمایشات خود، برخی از گام‌های اساسی تحقیقاتی پیش‌رو را اجرا کند. به‌طور کلی، خواننده پس از مطالعه پنج فصل اول کتاب، می‌تواند به‌تنهایی به تحلیل داده‌های NGS پردازد.

درحالی‌که عمده تجزیه و تحلیل NGS درحال‌حاضر با هم‌ردیفی شروع می‌شود، برنامه‌های دیگری وجود دارند که نیازمند سرهم‌بندی ژنوم هستند. این امر به‌ویژه درمورد ژنوم‌های کوچک‌تر و فناوری‌های NGS که طول خوانش بسیار طولانی تولید می‌کنند، مصداق داشته و مرسوم است. در آینده ممکن است مرز بین مرحله

1. Novocraft

هم‌ردیفی توالی و سرهم‌بندی ژنوم مشخص نباشد. در فصل ۶، دکتر تامی تجربیات خود را در مورد سرهم‌بندی توالی ژنوم با ارائه مقدمه‌ای در این خصوص به اشتراک می‌گذارد. در این فصل کتاب، او نه تنها نحوه سرهم‌بندی ژنوم را به خواننده نشان می‌دهد، بلکه روش ارزیابی کیفی آن را نیز آموزش می‌دهد. در چند فصل بعد، کتاب بر کاربردهای خاص روش‌های مبتنی بر NGS متمرکز شده است. این کتاب مجموعه‌ای از برنامه‌های کاربردی که به‌طور گسترده مورد استفاده قرار می‌گیرند را گزینش کرده است و نحوه پردازش داده‌ها برای هر برنامه را گام‌به‌گام آموزش می‌دهد. توالی‌یابی اگزوم به دلیل ملاحظات هزینه و عمق بیشتر پوشش، به شاخه مهمی از فناوری NGS تبدیل شده است. ما همچنین با استفاده از توالی‌یابی ترنسکریپتوم قادر هستیم درک بهتری از الگوی بیان ژن‌های هر سلول در لحظه داشته باشیم. یکی از جنبه‌های یکی دیگر از کاربردهای فناوری NGS، مطالعات متاژنومیکس است که سعی می‌کند برخلاف موارد کاربردی مورد بحث قبلی، پاسخ‌هایی در مورد محتوای ژنومی کل نمونه‌ها پیدا کند. سؤال مهم دیگر نحوه تعیین روابط بین ژنوتیپ‌ها و فنوتیپ‌هاست. همه این برنامه‌ها نیازمند رویکردهای متفاوتی برای اجرا هستند و انواع مختلفی از سؤالات را مطرح می‌کنند. با این حال، تکنیک‌های مورد استفاده در این حوزه‌ها را می‌توان در سایر روش‌های دیگر به کار گرفت. به عنوان مثال، تکنیک‌های مورد استفاده برای پردازش توالی‌یابی اگزوم می‌تواند در کار با سایر روش‌های توالی‌یابی هدفمند مفید باشد و تکنیک‌های استفاده‌شده برای یافتن تنوع در روش WGS نیز می‌توانند در مطالعات ترانسکریپتومیکس استفاده شوند. بنابراین، خواننده می‌تواند با درک مفاهیم اساسی مورد استفاده برای آنالیز انواع مختلف این مجموعه داده‌ها، از این کتاب بهره مند شود.

پروفسور نظر زکی^۱

رئیس تیم تحقیقات بیوانفورماتیک، هماهنگ‌کننده سیستم‌های هوشمند و توسعه نرم‌افزار

دانشکده فناوری اطلاعات، دانشگاه امارات متحده عربی

امارات متحده عربی، العین ۱۷۵۵۱

مقدمه

راز حیات در توالی DNA رمزگذاری شده است. از دهه ۱۹۷۰ تاکنون، بسیاری از مخترعان و مبتکران با توسعه فناوری‌های توالی‌یابی DNA ما را قادر ساخته‌اند تا بر فرایند پردردسر خوانش یک باز فائق بیاییم تا امروز بتوانیم توالی میلیون‌ها قطعه DNA را به راحتی خوانش کنیم. امروز، ما در دورانی زندگی می‌کنیم که فناوری‌های توالی‌یابی نسل جدید^۱ (NGS) تقریباً در دسترس هستند و توالی‌یاب‌های نسل سوم نیز تجاری شده‌اند. با این حال، به جرئت می‌توان گفت که مدیریت حجم عظیمی از داده‌های توالی‌یابی‌شده و استفاده از آن‌ها در مطالعات زیستی مختلف، چالش برانگیز است.

ایده نگارش این کتاب پس از تدریس و برگزاری کارگاه‌های مختلف NGS توسط من و همکارانم ایجاد شد. ما تصور کردیم که ارائه یک کتاب عملی کاربردی جامع در مورد NGS ایده خوبی خواهد بود تا افراد بیشتری بتوانند با چگونگی آنالیز داده‌های بیوانفورماتیک حاصل از این فناوری آشنا شوند. این کتاب شامل مباحث جامع در مورد چگونگی مدیریت داده‌های NGS از بررسی کیفیت توالی و سرهم‌بندی کردن خوانش‌ها برای یافتن چندشکلی تک‌نوکلئوتیدی (SNP) است. سایر موضوعات پیشرفته مانند سرهم کردن ژنوم، تعیین توالی اگزوم، آنالیز ترانسکریپتوم و متاژنومیکس نیز در این کتاب مطرح شده‌اند. به طور ویژه، آخرین فصل هم به کاربردهای داده‌های NGS اختصاص داده شده است تا کارایی این فناوری در نقشه‌یابی ژنتیکی و مطالعه ارتباطی جامع ژنومی (GWAS) را به خواننده نمایش دهد.

بسیاری از نوآموزانی که به یادگیری تجزیه و تحلیل داده‌های NGS نیاز دارند، با مشکلات مشترک و متداولی مواجه می‌شوند. این کتاب مطالب و تجربیات حاصل از آموزش بسیاری از فراگیران نخست را در کنار هم قرار داده که شامل منابع اضافی با هدف تقویت دانش خوانندگان در این زمینه است. اعتقاد داریم که این کتاب می‌تواند برای دانشجویان و محققان علوم زیستی بسیار مورد استفاده قرار گیرد. همچنین برای خوانندگانی که از قبل در تجزیه و تحلیل داده‌های مبتنی بر NGS مهارت دارند، این کتاب می‌تواند به عنوان یک مرجع کاربردی سودمند باشد.

قابل توجه خوانندگان گرامی:

مجموعه داده‌های همراه را می‌توان از آدرس زیر دانلود کرد:

<http://bioinfo.perdanauniversity.edu.my/infocenter/display/NPB/Index>