به نام خدا

تجزیه و تحلیل داده‌های توالی‌یابی نسل جدید

RICeST

تالیف:

زینکن وانگ

ترجمه:

سید ضیاالدین میرحسینی، شاهرخ قوتی، زهرا پزشکیان

سرشناسه	:	وانگ، شینکون، ۱۹۷۲- م. -Wang, Xinkun, 1972
عنوان و نام پدیدآور	:	تجزیه و تحلیل داده‌های توالی‌یابی نسل جدید/ تالیف زینکن وانگ؛ ترجمه سیدضیاءالدین میرحسینی، شاهرخ قوتی، زهرا پزشکیان.
مشخصات نشر	:	شیراز: مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، انتشارات، ۱۴۰۰.
مشخصات ظاهری	:	۲۷۷ ص.
شابک	:	978-622-97759-8-1
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	عنوان اصلی: Next-generation sequencing data analysis ,2016
یادداشت	:	کتابنامه:ص. ۲۵۷-۲۷۷.
موضوع	:	ژنگان‌شناسی
موضوع	:	Genomics
موضوع	:	ژنوم
موضوع	:	Genomes
شناسه افزوده	:	میرحسینی، ضیاءالدین، ۱۳۳۷-، مترجم
شناسه افزوده	:	قوتی، شاهرخ، ۱۳۵۸-، مترجم
شناسه افزوده	:	پزشکیان، زهرا، ۱۳۶۲-، مترجم
شناسه افزوده	:	ایران. وزارت علوم، تحقیقات و فناوری. مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری
شناسه افزوده	:	Iran. Ministery of Science, Research and Technology Regional Information Center for Science and Technology
رده بندی کنگره	:	QH۴۴۷
رده بندی دیویی	:	۸۶۲۹/۵۷۲
شماره کتابشناسی ملی	:	۸۴۳۴۵۱۹
اطلاعات رکورد کتابشناسی	:	فیپا

عنوان: تجزیه و تحلیل داده‌های توالی‌یابی نسل جدید
مترجمان: سید ضیاالدین میرحسینی، شاهرخ قوتی، زهرا پزشکیان
ناشر: انتشارات مرکز منطقه‌ای اطلاع رسانی علوم و فناوری
شابک: ۱-۸-۹۷۷۵۹-۶۲۲-۹۷۸
تیراژ: ۲۰۰ نسخه

حق چاپ برای مرکز منطقه‌ای اطلاع رسانی علوم و فناوری و دانشگاه گیلان محفوظ است.

اداره انتشارات مرکز منطقه‌ای اطلاع رسانی علوم و فناوری
شیراز بلوار جمهوری اسلامی خیابان جام جم Publication@ricest.ac.ir
تلفن: ۰۷۱۳۶۴۶۸۴۵۲

RICeST

فهرست

بخش ۱- مقدمه‌ای بر بیولوژی سلولی و مولکولی
۱. سیستم سلولی و کد حیات ۱
۱. ۱ چالش سلولی ۱
۱. ۲ سلول‌ها چگونه با چالش برخورد می‌کنند ۲
۱. ۳ مولکول‌ها در سلول‌ها ۳
۱. ۴ ساختارها یا فضاهای داخل سلولی ٤
۱. ۴. ۱. هسته ۴
۱. ۴. ۲. غشا سلولی ۶
۱. ۴. ۳ سیتوپلاسم ۶
۱. ۴. ۴ اندوزوم، لیزوزوم و پراکسیزوم ۷
۱. ۴. ۵ ریبوزوم ۸
۱. ۴. ۶ شبکه آندوپلاسمی ۹
۱. ۴. ۷ دستگاه گلژی ۱۰
۱. ۴. ۸ اسکلت سلولی ۱۰
۱. ۴. ۹ میتوکندری ۱۱
۱. ۴. ۱۰ کلروپلاست ۱۳
۱. ۵ سلول به عنوان یک سیستم ١٤
۱. ۵. ۱. سیستم سلولی ۱۴
۱. ۵. ۲ زیست شناسی سیستم سلول ۱۵
۱. ۵. ۳. چگونگی مطالعه سیستم سلولی ۱۶
۲. توالی DNA: اساس ژنوم ۱۸
۲. ۱. مارپیچ دو رشته‌ای DNA و توالی باز ۱۸
۲. ۲ چگونه مولکول‌های DNA همانند سازی می‌کنند و فرم اصلی را حفظ می‌نمایند ۱۹
۲. ۳ چگونه اطلاعات ژنتیکی ذخیره شده در DNA به پروتئین انتقال می‌یابند ۲۱
۲. ۴ چشم انداز ژنومی ۲۳
۲. ۴. ۱. ژنوم حداقلی ۲۳
۲. ۴. ۲ اندازه‌های ژنومی ۲۳
۲. ۴. ۳ مناطق کد کننده پروتئین در ژنوم ۲۴
۲. ۴. ۴ عناصر ژنومی غیر کد کننده ۲۵
۲. ۵ بسته بندی DNA، دسترسی به توالی و تعاملات DNA- پروتئین ۲۸

RICeST

۲. ۵. ۱ بسته بندی DNA ۲۸
۲. ۵. ۲ دسترسی به توالی ۲۸
۲. ۵. ۳ تعاملات پروتئین- DNA ۲۹
۲. ۶ جهش و پلی مورفیسم توالیDNA ۳۰
۲. ۷ تکامل ژنوم ۳۳
۲. ۸ اپی ژنوم و متیلاسیونDNA ۳٤
۲. ۹ توالی‌یابی ژنوم و خطر بیماری ۳٦
۲. ۹. ۱ بیماری‌های مندلی (تک ژنی) ۳۶
۲. ۹. ۲ بیماری‌های پیچیده که ژن‌های متعدد را شامل می‌شوند ۳۷
۲. ۹. ۳ بیماری‌هایی که ناشی از ناپایداری ژنوم است ۳۷
۲. ۹. ۴ بیماریهای اپی ژنومیک/اپی ژنتیکی ۳۸
۳. RNA: توالی رونویسی شده ۴۰
۳. ۱ RNA به عنوان پیام‌رسان ٤۰
۳. ۲ ساختار مولکولی RNA ٤۰
۳. ۳. ۱ الگوی DNA ۴۲
۳. ۳. ۲ رونویسی ژن‌های پروکاریوتی ۴۳
۳. ۳. ۳ رونویسی اولیه پیش‌ساز mRNA ژن‌های یوکاریوتی ۴۵
۳. ۳. ۴ بلوغ mRNA از پیش‌ساز mRNA ۴۷
۳. ۳. ۵ انتقال و محل‌یابی ۵۰
۳. ۳. ۶ پایداری و زوال ۵۰
۳. ۳. ۷ مراحل اصلی تنظیم سطح رونویسی mRNA ۵۲
۳. ۴ RNA بیشتر از یک پیام رسان است ٥۳
۳. ۴. ۱ رایبوزایم ۵۴
۳. ۴. ۲ snRNA وsnoRNA ۵۵
۳. ۴. ۳ RNA برای همانندسازی تلومر ۵۵
۳. ۴. ۴ RNAi وRNAهای غیرکدکننده کوچک ۵۶
۳. ۴. ۵ RNAهای غیر کد کننده طویل ۶۰
۳. ۴. ۶ دیگر RNAهای غیر کد کننده ۶۱
۳. ۵ چشم انداز ترانسکریپتی سلولی ٦۱
بخش ۲- مقدمه‌ای بر توالی‌یابی نسل جدید (NGS) و آنالیز داده‌های NGS
۴. تکنولوژی‌های توالی‌یابی نسل جدید (NGS): ورودی‌ها و خروجی‌ها ۶۴
۴. ۱ نحوه توالی‌یابی DNA: از نسل اول تا نسل جدید ٦٤

RICeST

۴. ۲ گردش کار تجربی یک توالی‌یابی نسل جدید معمولی ۶۸
۴. ۳ ورودی‌ها و خروجی‌های انواع پلت فرم‌های NGS ۷۱
۴. ۳. ۱ توالی‌یابی خاتمه دهنده با رنگ برگشت‌پذیر ایلومینا ۷۲
۴. ۳. ۲ توالی‌یابی نیمه‌هادی آیون تورنت ۷۸
۴. ۳. ۳ توالی‌یابی زمان واقعی تک مولکولی پسیفیک بایوساینسز (SMRT) ۸۰
۴. ۴ اریبی‌ها و سایر عوامل اشتباه که ممکن است دقت داده‌های NGS را تحت تاثیر قرار دهند ۸۱
۴. ۴. ۱ اریبی‌ها در ساخت کتابخانه ۸۳
۴. ۴. ۲ اریبی‌ها و عوامل دیگر در توالی‌یابی ۸۴
۴. ۵ برنامه‌های کاربردی عمده NGS ۸۵
۴. ۵. ۱ تعیین مشخصات ترانسکریپتوم و تشخیص تنوع پیرایشی (RNA-Seq) ۸۵
۴. ۵. ۲ جهش ژنتیکی و کشف تغییرات ۸۵
۴. ۵. ۳ اسمبل کردن ژنوم به روش از نو (د نوو) ۸۶
۴. ۵. ۴ تجزیه و تحلیل متقابل پروتئین –DNA (ChIP-Seq) ۸۶
۴. ۵. ۵ مطالعات اپی ژنومیک و متیلاسیون DNA (Methyl-Seq) ۸۶
۴. ۵. ۶ متاژنومیکس ۸۷
۵. آنالیز داده‌های توالی‌یابی نسل جدید (NGS) در مراحل اولیه: مراحل مشترک ۸۸
۵. ۱ فراخوانی باز، فرمت فایل FASTQ، و نمره کیفیت باز ۸۹
۵. ۲ کنترل کیفیت و پیش پردازش داده‌های NGS ۹۳
۵. ۳ مکان‌یابی خوانش‌ها ۹۵
۵. ۳. ۱ روش‌ها و الگوریتم‌های مکان‌یابی ۹۵
۵. ۳. ۲ انتخاب الگوریتم‌های مکان‌یابی و توالی‌های ژنوم مرجع ۹۸
۵. ۳. ۳ SAM / BAM به عنوان فرمت استاندارد مکان‌یابی ۹۹
۵. ۳. ۴ بررسی و عملیات فایل مکان‌یابی ۱۰۲
۶. نیازمندی‌های محاسباتی برای مدیریت و تجزیه و تحلیل داده‌ها در توالی‌یابی نسل جدید (NGS) ۱۰۶
۶. ۱ ذخیره سازی، انتقال و به اشتراک گذاشتن داده‌های NGS ۱۰۶
۶. ۲ قدرت محاسباتی مورد نیاز برای تجزیه و تحلیل داده NGS ۱۰۹
۶. ۳ نیازهای نرم افزاری برای تجزیه و تحلیل داده NGS ۱۱۰
۶. ۴ مهارت‌های بیوانفورماتیکی مورد نیاز برای تجزیه و تحلیل داده NGS ۱۱۳
بخش ۳- تجزیه و تحلیل اختصاصی کاربردی داده‌های NGS

RICeST

۷. ترانسکریپتومیکس توسط توالی‌یابی RNA ۱۱۷
۷. ۱ اصول RNA-Seq ۱۱۷
۷. ۲ طراحی آزمایش ۱۱۸
۷. ۲. ۱ طرح فاکتوریل ۱۱۸
۷. ۲. ۲ تکرار و تصادفی بودن ۱۱۹
۷. ۲. ۳ آماده سازی نمونه ۱۱۹
۷. ۲. ۴ استراتژی توالی‌یابی ۱۲۱
۷. ۳ تجزیه و تحلیل داده‌های RNA-Seq ۱۲۲
۷. ۳. ۱ کنترل کیفیت داده‌ها و مکان‌یابی خوانش‌ها ۱۲۲
۷. ۳. ۲ نرمال سازی داده‌های RNA-Seq ۱۲۵
۷. ۳. ۳ شناسایی ژن ها با بیان متفاوت ۱۲۸
۷. ۳. ۴ آنالیز پیرایشی متفاوت ۱۳۱
۷. ۳. ۵ تجسم داده‌های RNA-Seq ۱۳۲
۷. ۳. ۶ آنالیز عملکردی ژن‌های شناسایی شده ۱۳۳
۷. ۴ RNA-Seq به عنوان ابزار کشف ۱۳۳
۸. توالی‌یابی RNA کوچک ۱۳۵
۸. ۱ تولید داده‌ها و پردازش‌های بالادستی RNA کوچک توسط توالی‌یابی نسل جدید (NGS) ۱۳۶
۸. ۱. ۱ تولید داده‌ها ۱۳۶
۸. ۱. ۲ پیش پردازش ۱۳۸
۸. ۱. ۳ مکان‌یابی ۱۳۹
۸. ۱. ۴ شناسایی گونه‌های RNA کوچک شناخته شده و احتمالی ۱۴۱
۸. ۱. ۵ نرمال سازی ۱۴۱
۸. ۲ شناسایی RNAهای کوچک با بیان متفاوت ۱۴۲
۸. ۳ آنالیز عملکردی RNAهای کوچک شناسایی شده ۱۴۲
۹. تعیین ژنوتیپ و کشف تغییرات ژنوم توسط توالی‌یابی مجدد کل ژنوم ۱۴۴
۹. ۱ پیش پردازش، مکان‌یابی، همردیفی مجدد، و کالیبراسیون مجدد داده‌ها ۱۴۶
۹. ۲ تنوع تک نوکلئوتیدی (SNV) و فراخوانی ایندل ۱۴۷
۹. ۲. ۱ فراخوانی SNV ۱۴۷
۹. ۲. ۲ شناسایی جهش‌ها به روش از نو یا د نوو ۱۴۹
۹. ۲. ۳ فراخوانی کردن ایندل ۱۵۱
۹. ۲. ۴ فراخوانی تنوع داده‌های RNA-Seq ۱۵۱

RICeST

۹. ۲. ۵ فایل فرمت فراخوانی تنوع (VCF) ۱۵۲
۹. ۲. ۶ ارزیابی نتایج VCF ۱۵۴
۹. ۳ فراخوانی تنوع ساختاری (SV) ۱۵۵
۹. ۳. ۱ فراخوانی SV مبتنی بر خوانش جفتی ۱۵۵
۹. ۳. ۲ تعیین نقطه شکست ۱۵۶
۹. ۳. ۳ تشخیص SV مبتنی بر اسمبل به روش از نو ۱۵۷
۹. ۳. ۴ تشخیص CNV ۱۵۷
۹. ۳. ۵ آنالیز یکپارچه SV ۱۵۸
۹. ۴ حاشیه نویسی تنوع‌های فراخوان شده ۱۵۸
۹. ۵ آزمودن ارتباط تنوع با بیماری‌ها یا صفات ۱۵۹
۱۰. اسمبل کردن ژنوم به روش از نو از خوانش‌های توالی‌یابی نسل جدید (NGS) **۱۶۱**
۱۰. ۱ عوامل ژنتیکی و استراتژی‌های توالی‌یابی برای اسمبل کردن به روش از نو ۱۶۲
۱۰. ۱. ۱ عوامل ژنتیکی مؤثر بر اسمبل کردن به روش از نو ۱۶۲
۱۰. ۱. ۲ استراتژی‌های توالی‌یابی برای اسمبل کردن به روش از نو ۱۶۳
۱۰. ۲. اسمبل کردن کانتیگ ها ۱۶۵
۱۰. ۲. ۱ پردازش داده‌های توالی‌یابی، تصحیح خطا و ارزیابی خصوصیات ژنوم ۱۶۵
۱۰. ۲. ۲ الگوریتم‌های اسمبلی کانتیگ ۱۶۷
۱۰. ۳ اسکافولدینگ ۱۷۰
۱۰. ۴ ارزیابی کیفیت اسمبلی ۱۷۱
۱۰. ۵ بسته شدن شکاف ۱۷۲
۱۰. ۶ محدودیت ها و توسعه آینده ۱۷۳
۱۱. مکان یابی برهمکنش پروتئین و DNA با ChIP-Seq **۱۷۵**
۱۱. ۱ اصول ChIP-Seq ۱۷۵
۱۱. ۲ طراحی آزمایش ۱۷۷
۱۱. ۲. ۱ کنترل آزمایش ۱۷۷
۱۱. ۲. ۲ عمق توالی‌یابی ۱۷۸
۱۱. ۲. ۳ تکرار ۱۷۹
۱۱. ۳ مکان یابی خوانش، فراخوانی پیک (اوج) و مجسم‌سازی اوج ۱۷۹
۱۱. ۳. ۱ کنترل کیفیت داده‌ها و مکان یابی خوانش ۱۷۹
۱۱. ۳. ۲ فراخوانی اوج ۱۸۳
۱۱. ۳. ۳ مجسم سازی اوج ۱۹۱
۱۱. ۴ آنالیز اتصال متفاوت ۱۹۱

RICeST

۱۱. ۵ تجزیه و تحلیل عملکردی ۱۹۴
۱۱. ۶ تجزیه و تحلیل موتیف ۱۹۵
۱۱. ۷ تجزیه و تحلیل یکپارچه داده‌های ChIP-Seq ۱۹۶
۱۲. تجزیه و تحلیل اپی ژنومیکس و متیلاسیون DNA توسط توالی‌یابی نسل جدید (NGS) ۱۹۸
۱۲. ۱ استراتژی‌های توالی‌یابی متیلاسیون DNA ۱۹۹
۱۲. ۱. ۱ توالی‌یابی بی‌سولفیت کل ژنوم (WGBS) ۱۹۹
۱۲. ۱. ۲ توالی‌یابی با کاهش بازنمایی بی‌سولفیت (RRBS) ۲۰۱
۱۲. ۱. ۳ توالی‌یابی متیلاسیون بر اساس غنی سازی DNA متیله ۲۰۱
۱۲. ۱. ۴ تمایز متیلاسیون سیتوزین از محصولات دمتیلاسیون در توالی‌یابی بی‌سولفیت ۲۰۲
۱۲. ۲ تجزیه و تحلیل داده‌های توالی‌یابی متیلاسیون DNA ۲۰۳
۱۲. ۲. ۱ کنترل کیفیت و پیش پردازش ۲۰۳
۱۲. ۲. ۲ مکان یابی خوانش ۲۰۳
۱۲. ۲. ۳ تعیین مقدار متیلاسیون DNA ۲۰۶
۱۲. ۲. ۴ تجسم داده‌های متیلاسیون DNA ۲۰۸
۱۲. ۳ تشخیص سیتوزین‌ها یا مناطقی که به صورت متفاوت متیله شده ۲۱۰
۱۲. ۴ تأیید، اعتبارسنجی و تفسیر داده‌ها ۲۱۱
۱۳. تجزیه و تحلیل متاژنوم توسط توالی‌یابی نسل جدید (NGS) ۲۱۲
۱۳. ۱ طراحی آزمایش و تهیه نمونه ۲۱۳
۱۳. ۱. ۱ مجموعه نمونه متاژنوم ۲۱۵
۱۳. ۱. ۲ پردازش نمونه متاژنوم ۲۱۵
۱۳. ۲ روش‌های توالی‌یابی ۲۱۶
۱۳. ۳ بررسی کلی تجزیه و تحلیل داده‌های توالی‌یابی متاژنوم شات گان کل ژنوم (WGS) ۲۱۷
۱۳. ۴ کنترل کیفیت و پیش پردازش داده‌های توالی‌یابی ۲۱۹
۱۳. ۵ خصوصیات تاکسونومیک یک جامعه میکروبی ۲۲۰
۱۳. ۵. ۱ اسمبل کردن متاژنوم ۲۲۰
۱۳. ۵. ۲ گذاشتن توالی در پیاله ۲۲۱
۱۳. ۵. ۳ فراخوانی قاب‌های خوانش باز (ORFها) و سایر عناصر ژنومی از توالی‌های متاژنومی ... ۲۲۳
۱۳. ۵. ۴ تجزیه و تحلیل نشانگر ژن فیلوژنتیکی ۲۲۴
۱۳. ۶ خصوصیات عملکردی یک جامعه میکروبی ۲۲۵
۱۳. ۶. ۱ حاشیه نویسی عملکرد ژن ۲۲۵
۱۳. ۶. ۲ بازسازی مسیر متابولیک ۲۲۶
۱۳. ۷ آنالیز متاژنومی مقایسه‌ای ۲۲۶

RICeST

۱۳. ۷. ۱ نرمال سازی داده‌های توالی‌یابی متاژنوم ۲۲۷
۱۳. ۷. ۲ شناسایی گونه‌های فراوان یا واحدهای تاکسونومیکی عملیاتی (OTU ها) مختلف..... ۲۲۷
۱۳. ۸ خط مشی تجزیه و تحلیل داده‌های متاژنومیکس..... ۲۲۸
۱۳. ۹ مخازن داده متاژنومیکس ۲۲۸
بخش ۴- تغییر چشم انداز فن‌آوری‌های توالی یابی نسل جدید و تجزیه و تحلیل داده‌ها
۱۴. آینده توالی‌یابی نسل جدید (NGS) چیست؟ ۲۳۱
۱۴. ۱ منظره در حال تغییر توالی‌یابی نسل جدید (NGS)..... ۲۳۱
۱۴. ۲ تکامل سریع و رشد ابزارهای بیوانفورماتیک برای تجزیه و تحلیل داده‌های توالی‌یابی با توان بالا ۲۳۳
۱۴. ۳ استاندارد سازی و کارآمد سازی خط مشی تحلیلی NGS..... ۲۳۵
۱۴. ۴ محاسبات موازی ۲۳۶
۱۴ .۵ محاسبات ابری ۲۳۷
پیوست الف: انواع فایل‌های رایج مورد استفاده در تجزیه و تحلیل داده‌های توالی‌یابی نسل جدید (NGS)..... ۲۴۰
پیوست ب: واژه نامه ۲۴۴
منابع ۲۵۷

RICeST

RICeST

بخش ۱

مقدمه‌ای بر بیولوژی سلولی و مولکولی

RICeST

RICeST

۱

سیستم سلولی و کد حیات

۱.۱ چالش سلولی

اگرچه یک سلول جسم ریزی با قطر کمتر از ۵۰ میکرومتر است، اما در مقایسه با هر نوع سیستم دست ساز انسانی به طور شگفت انگیزی کار می‌کند. علاوه بر این، سلول با استفاده از اطلاعات کد گذاری شده در DNA خود، حیات خویش را تداوم می‌بخشد. اگر شما تا به حال فکر طراحی یک سیستم مصنوعی را داشته‌اید که این نوع از پیچیدگی را نشان دهد، می‌دانید که چنین سیستمی نیاز به غلبه بر چالش‌های غیرقابل حلی دارد. یک سلول سیستم داخلی پیچیده‌ای دارد که حاوی انواع بسیاری از مولکول‌ها و بخش‌ها است. یک سلول، برای حفظ این سیستم، نیاز به انجام طیف گسترده‌ای از وظایف در محیطی که به طور مداوم در حال تغییر است، دارد. اساسی‌ترین این وظایف حفظ نظم داخلی سلول، جلوگیری از عملکرد نادرست و تجزیه شدن سیستم، و تکثیر یا حتی توسعه این سیستم است.

انرژی برای حفظ نظم داخلی سیستم سلولی مورد نیاز است. بدون انرژی ثابت ورودی، بی‌نظمی[1] سیستم همان‌طور که در قانون دوم ترمودینامیک آمده است، به تدریج افزایش می‌یابد، و در نهایت منجر به تخریب سیستم می‌شود. علاوه بر انرژی، از آنجایی که ساختار داخلی یک سلول، ساختاری پویا است و به تغییرات مداوم در شرایط محیطی پاسخ می‌دهد، مواد "ساختمانی" خام نیز برای تجدید قطعات داخلی یا ساخت قطعات جدید مورد نیاز است. بنابراین، برای حفظ تعادل در داخل سلول و همچنین حفظ تعادل با محیط، نیاز به ورود مداوم انرژی و مواد خام و دفع مواد زائد است. هدایت چگونگی به دست آوردن انرژی مورد نیاز و مواد اولیه برای بقا و تداوم سلول، بر عهده اطلاعات کد گذاری شده در توالی DNA آن است.

1. Entropy

RICeST