

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



انتشارات  
۸۲۵

## کاربرد نرم افزار R در آنالیز داده های ژنتیکی

دکتر محمد تیموریان

دکتر محمدمهدی شریعتی

عضو هیئت علمی دانشگاه فردوسی مشهد

سرشناسه: تیموریان، محمد، ۱۳۶۱ -  
 عنوان و نام پدیدآور: کاربرد نرم افزار R در آنالیز داده های ژنتیکی / محمد تیموریان، محمدمهدی شریعتی؛ ویراستار  
 ادبی هانیه اسدیورفعال مشهد.  
 مشخصات نشر: مشهد: دانشگاه فردوسی مشهد، انتشارات، ۱۴۰۱.  
 مشخصات ظاهری: ۱۹۲ ص: جدول، نمودار.  
 فروست: انتشارات دانشگاه فردوسی مشهد؛ ۸۲۵.  
 شابک: ISBN: 978-964-386-517-7  
 وضعیت فهرست نویسی: فیبا.  
 یادداشت: کتاینامه: ص. [۱۸۷] - ۱۹۰. نمایه.  
 موضوع: آر (زبان برنامه نویسی کامپیوتر)  
 ژنتیک -- روش های آماری  
 شریعتی، محمدمهدی، ۱۳۵۲ -  
 شناسه افزوده: دانشگاه فردوسی مشهد، انتشارات.  
 شناسه افزوده: Q۸۲۷۶/۴۵  
 رده بندی کنگره: ۵۱۹/۵۰۲۸۵۵۱۳۳  
 رده بندی دیویی: ۸۷۴۶۷۹۳  
 شماره کتابشناسی ملی:

R (Computer program language )  
 Genetics -- Statistical methods

## کاربرد نرم افزار R در آنالیز داده های ژنتیکی

پدیدآورندگان: دکتر محمد تیموریان؛ دکتر محمدمهدی شریعتی  
 ویراستار ادبی: هانیه اسدیور فعال مشهد  
 مشخصات: وزیری، ۱۰۰ نسخه، چاپ اول، تابستان ۱۴۰۱  
 چاپ و صحافی: چاپخانه دقت  
 بها: ۹۹۰/۰۰۰ ریال  
 حق چاپ برای انتشارات دانشگاه فردوسی مشهد محفوظ است.



انتشارات  
 ۸۲۵

### 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، بن بست  
 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](mailto:press@um.ac.ir)

تقدیم به:

همسر و فرزندانم، سنا و سایدا

محمد تیموریان

تقدیم به:

شادروان دکتر فریدون افتخار شاهرودی

محمد مهدی شریعتی

press.um.ac.ir

**press.um.ac.ir**

## فهرست مطالب

پیشگفتار.....	۱۲
<b>فصل ۱. کلیات نرم افزار R.....</b>	<b>۱۳</b>
۱-۱ نصب نرم افزار در ویندوز.....	۱۳
۲-۱ نصب نرم افزار در لینوکس.....	۱۳
۳-۱ نصب کتابخانه.....	۱۴
۱-۳-۱ مخزن Bioconductor.....	۱۴
۲-۳-۱ مخزن گیت هاب.....	۱۵
۳-۳-۱ کتابخانه Rcmdr.....	۱۵
۴-۱ نکات کلی.....	۱۵
۵-۱ فراخوانی داده ها.....	۱۷
۶-۱ ذخیره داده ها.....	۲۰
۷-۱ مشاهده و بررسی داده ها.....	۲۱
<b>فصل ۲. انواع داده ها در R.....</b>	<b>۲۳</b>
۱-۲ بردار.....	۲۳
۲-۲ ماتریس.....	۲۴
۱-۲-۲ وارون ماتریس.....	۲۷
۳-۲ آرایه.....	۲۸
۴-۲ لیست.....	۲۸
۵-۲ قالب جدولی داده ها.....	۲۸
۶-۲ نمونه گیری.....	۳۰
۱-۶-۲ کتابخانه dplyr.....	۳۲

۳۳	فصل ۳. توالی‌های ژنتیکی در R .....
۳۳	۱-۳ بررسی کلی توالی‌ها .....
۳۶	۲-۳ بررسی توالی‌ها با کتابخانه Biostrings .....
۳۷	۱-۲-۳ پوشش موقتی مناطق خاص .....
۳۹	فصل ۴. توابع ریاضی و آماری .....
۳۹	۱-۴ عملیات و توابع ریاضی در R .....
۴۰	۱-۱-۴ تبدیل تاریخ به فصل .....
۴۰	۲-۴ حل معادله .....
۴۱	۳-۴ مشتق و انتگرال .....
۴۱	۴-۴ توابع مهم آماری .....
۴۳	۵-۴ جدول توافقی .....
۴۴	۶-۴ توزیع‌های تصادفی .....
۴۴	۱-۶-۴ تابع چگالی .....
۴۴	۲-۶-۴ تابع توزیع .....
۴۵	۳-۶-۴ توابع صدکی و چندکی .....
۴۵	۴-۶-۴ تولید اعداد تصادفی .....
۴۵	۷-۴ نمودارهای آماری .....
۴۵	۱-۷-۴ نمودار پراکنش .....
۴۷	۲-۷-۴ نمودار دایره‌ای و میله‌ای .....
۴۷	۳-۷-۴ نمودار جعبه‌ای .....
۴۸	۴-۷-۴ نمودار هیستوگرام .....
۴۸	۵-۷-۴ نمودار حرارتی .....
۴۸	۶-۷-۴ نمودار MA .....
۴۹	۷-۷-۴ ذخیره نمودارها .....
۴۹	۸-۴ نوشتن تابع .....
۵۰	۱-۸-۴ حلقه تکرار و شرط در توابع .....

۹-۴	اعمال کردن توابع	۵۱
۱۰-۴	فاصله و کاهش ابعاد	۵۲
۱۱-۴	آنالیز مؤلفه‌های اصلی	۵۲
۱۲-۴	خوشه‌بندی چندگانه	۵۴
۱۳-۴	خوشه‌بندی سلسله‌مراتبی	۵۴
۱-۱۳-۴	نمودار حرارتی داده‌های ژنومی	۵۶

## فصل ۵. آزمون‌های آماری و استنباطی

۱-۵	آزمون کولموگروف اسمیرنوف	۵۷
۲-۵	آزمون تک‌نمونه‌ای	۵۷
۳-۵	آزمون مقایسه‌ی دونمونه‌ای مشاهدات مستقل	۵۸
۴-۵	آزمون دونمونه‌ای مشاهدات زوجی	۵۹
۵-۵	آزمون نیکویی برازش کای مربع	۵۹
۶-۵	آزمون استقلال کای مربع	۶۰
۷-۵	هم‌بستگی	۶۱
۸-۵	رگرسیون و مدل‌های خطی	۶۲
۹-۵	تجزیه‌ی واریانس	۶۳
۱۰-۵	طرح آزمایشات	۶۶
۱-۱۰-۵	آزمون‌های تعقیبی	۶۶
۱۱-۵	تصحیح معنی‌داری آزمون‌های چندگانه	۶۷
۱۲-۵	مفهوم p-value در مقایسات میانگین	۶۷

## فصل ۶. برآورد اثرات

۱-۶	ماتریس ضرایب	۶۹
۲-۶	برآورد اثرات ثابت با روش حداقل مربعات	۷۰
۱-۲-۶	برآورد اثرات ثابت با روش ماتریسی	۷۱
۲-۲-۶	روش تجزیه‌ی چالسکی و تکرار	۷۲

۶-۳ تشکیل ماتریس روابط خویشاوندی ..... ۷۳

۶-۴ پیش‌بینی اثرات تصادفی در مدل‌های مختلط با BLUP ..... ۷۳

**فصل ۷. آنالیزهای ارتباط ژنی** ..... ۷۵

۷-۱ پردازش کلی داده‌های QTL ..... ۷۵

۷-۲ بررسی مدل‌های مطالعات ارتباط ژنی ..... ۷۷

۷-۳ تغییر فرمت ژنوتیپ داده‌های نشانگری SNP ..... ۷۹

۷-۴ انتخاب به کمک نشانگر ..... ۸۰

۷-۴-۱ برآورد اثرات نشانگری ..... ۸۰

۷-۴-۲ پیش‌بینی اثرات تصادفی دام در روش انتخاب به کمک نشانگر ..... ۸۱

**فصل ۸. آنالیزهای GWAS در R** ..... ۸۳

۸-۱ فراخوانی داده‌ها ..... ۸۳

۸-۲ پایگاه داده SQL ..... ۸۴

۸-۳ آماده‌سازی و کنترل کیفیت داده‌های GWAS ..... ۸۶

۸-۳-۱ کنترل کیفیت نشانگرها ..... ۸۶

۸-۳-۲ کنترل کیفیت نمونه‌ها ..... ۸۸

۸-۴ آنالیزهای تک‌نشانگری در GWAS ..... ۹۰

۸-۵ آنالیز چندگانه اثرات نشانگری ..... ۹۳

۸-۵-۱ تصحیح بنفرونی معنی‌داری در آنالیزهای چندگانه ..... ۹۳

۸-۵-۲ برآورد هم‌زمان اثرات نشانگری با روش انقباضی SNP-BLUP ..... ۹۵

۸-۶ نمودار منتهن داده‌های GWAS ..... ۹۸

**فصل ۹. پیش‌بینی ژنومی در R** ..... ۹۹

۹-۱ آنالیزهای انتخاب ژنومی به روش BLUP ..... ۹۹

۹-۲ پیش‌بینی ژنومی ..... ۱۰۱



۳-۹	پیش‌بینی با GBLUP	۱۰۲
۴-۹	نشانه‌های انتخاب	۱۰۳
۵-۹	محاسبه عدم تعادل لینکاژی	۱۰۶
۶-۹	رسم نمودار درختی فواصل ژنتیکی	۱۰۶
۷-۹	فواصل ژنتیکی با استفاده از ماتریس روابط ژنومی G	۱۰۷
۸-۹	تحلیل مؤلفه‌های اصلی روابط ژنومی	۱۰۷

### فصل ۱۰. آنالیزهای بیان ژن در R- داده‌های آرایه‌ای

۱-۱۰	فراخوانی داده‌های آرایه‌ای	۱۱۰
۲-۱۰	کنترل کیفیت داده‌های آرایه‌ای	۱۱۰
۳-۱۰	پیش‌پردازش داده‌های آرایه‌ای	۱۱۳
۴-۱۰	آنالیزهای بیان افتراقی ژنی داده‌های آرایه‌ای	۱۱۵
۵-۱۰	آنالیز چندگانه ریزآرایه‌ها	۱۱۸

### فصل ۱۱. آنالیزهای بیان ژن در R - داده‌های توالی

۱-۱۱	فراخوانی داده‌های SRA	۱۲۱
۲-۱۱	فراخوانی داده‌های fastq	۱۲۲
۳-۱۱	دانلود فایل‌های ژنوم مرجع و آدرس‌دهی ژن‌ها (حاشیه‌نویسی)	۱۲۲
۴-۱۱	شاخص‌سازی ژنوم مرجع	۱۲۳
۵-۱۱	کنترل کیفیت داده‌های توالی RNA با نرم‌افزار fastQC	۱۲۴
۶-۱۱	کنترل کیفیت داده‌های توالی RNA	۱۲۴
۷-۱۱	پیش‌پردازش داده‌های توالی RNA	۱۲۵
۸-۱۱	پیش‌پردازش داده‌های توالی RNA با نرم‌افزار trimmomatic	۱۲۶
۹-۱۱	هم‌ردیفی داده‌ها با ژنوم مرجع	۱۲۷
۱۰-۱۱	هم‌ردیفی با کتابخانه Rsubread	۱۲۸
۱۱-۱۱	بررسی فایل‌های هم‌ردیف‌شده با کتابخانه GenomicAlignments	۱۲۹
۱۲-۱۱	شمارش تعداد خوانش‌ها با featureCounts	۱۳۰

**فصل ۱۲. پردازش داده‌های شمارش** ..... ۱۳۳

- ۱-۱۲ پیش‌پردازش داده‌های شمارش ..... ۱۳۳
- ۲-۱۲ تبدیل فایل شمارش به فایل ورودی DESeq2 ..... ۱۳۶
- ۳-۱۲ نرمال‌سازی داده‌های شمارش ..... ۱۳۶
- ۴-۱۲ نرمال‌سازی داده‌های شمارش برای اریبی حاصل از ترکیبات ..... ۱۴۱
- ۵-۱۲ آنالیز مؤلفه‌های اصلی ..... ۱۴۳
- ۶-۱۲ نمودار مقیاس‌گذاری چندبعدی ..... ۱۴۴
- ۷-۱۲ خوشه‌بندی سلسله‌مراتبی ..... ۱۴۷
- ۸-۱۲ تولید ماتریس ضرایب در آنالیزهای افتراقی ژن ..... ۱۴۹

**فصل ۱۳. آنالیز افتراقی بیان ژن** ..... ۱۵۱

- ۱-۱۳ آنالیز افتراقی بیان ژن داده‌های شمارش ..... ۱۵۱
- ۲-۱۳ افزودن حاشیه‌ها به نتایج آزمون افتراقی ژن‌ها با کتابخانه `org.Mm.eg.db` ..... ۱۵۷
- ۳-۱۳ افزودن حاشیه‌ها به نتایج آزمون افتراقی ژن‌ها با کتابخانه `biomaRt` ..... ۱۵۸
- ۴-۱۳ بررسی موقعیت‌های ژنومی از طریق کتابخانه‌های پایگاه اطلاعات رونوشت ژنی ..... ۱۶۰
- ۵-۱۳ نمودارهای آزمون افتراقی ژن‌ها ..... ۱۶۳
- ۱-۵-۱۳ رسم نمودارهای افتراق ژنی با کتابخانه `ggplot2` ..... ۱۶۶
- ۲-۵-۱۳ نمودار نواری بیان ژن ..... ۱۶۷
- ۶-۱۳ تولید فایل قابل‌بارگذاری از نتایج آزمون افتراقی در مرورگرها ..... ۱۶۹
- ۷-۱۳ ساخت نمودارها با کتابخانه `ggbio` ..... ۱۷۲

**فصل ۱۴. آزمون‌های تعقیبی بیان افتراقی ژن** ..... ۱۷۵

- ۱-۱۴ آزمون‌های مجموعه ژنی ..... ۱۷۵
- ۱-۱-۱۴ آزمون مجموعه ژنی رقابتی با `goana` ..... ۱۷۶
- ۲-۱-۱۴ آزمون مجموعه ژنی رقابتی با `GOseq` ..... ۱۷۷
- ۳-۱-۱۴ آزمون مجموعه ژنی جامع با `ROAST` ..... ۱۷۸

۱۸۰	..... ۲-۱۴ آنالیزهای ماهیت ژنی (هستی‌شناسی ژن)
۱۸۰	..... ۱-۲-۱۴ ماهیت‌شناسی ژن با CAMERA
۱۸۱	..... ۳-۱۴ آنالیزهای غنی‌سازی
۱۸۲	..... ۱-۳-۱۴ آنالیزهای غنی‌سازی مجموعه ژنی با fgsea
۱۸۴	..... ۲-۳-۱۴ آنالیزهای غنی‌سازی مسیر KEGG
۱۸۵	..... ۴-۱۴ نمودارهای آزمون مجموعه ژنی
۱۸۷	..... منابع
۱۹۱	..... نمایه

## پیشگفتار

ستایش خدای را که باران رحمت بی حسابش همه را رسیده و خوان نعمت بی دریغش همه جا کشیده و درود بر رسول گرامی اش که تاریکی جهل و نادانی را با نور جمال معرفت خود زائل کرد. در دهه‌های اخیر، مطالعات زیستی و ژنتیکی با سرعت بسیار زیادی در حال انجام بوده است. با پیشرفت‌های زیست‌شناسی مولکولی و ژنتیک و همچنین گسترش توالی‌یابی ژنوم‌های مختلف، تجزیه و تحلیل داده‌های ژنتیکی و فهم مسائل مرتبط با آن چالش بزرگی را برای زیست‌شناسان ایجاد کرده است. بررسی و آنالیز توالی‌های ژنومی گونه‌های مختلف و داده‌های نسل سوم، نیازمند نرم‌افزارهای قدرتمند آماری و زبان‌های مختلف برنامه‌نویسی است. نرم‌افزار R محیط بسیار مناسبی برای محاسبات و آنالیزهای آماری در بسیاری از رشته‌هاست و به دلیل رایگان بودن و نصب بر روی اکثر سیستم‌ها در سال‌های اخیر توجه کاربران زیادی را به خود جلب کرده است. همچنین امکان نصب بسته‌ها یا کتابخانه‌های متنوع در رشته‌های مختلف، قدرت زیادی به این نرم‌افزار داده است. با توجه به پیشرفت‌های صورت گرفته در علم ژنتیک و افزایش داده‌های مختلف این رشته در کشور عزیزمان، ایران، بر آن شدیم که آنالیز داده‌های ژنتیکی را به صورت مفید و مختصر با استفاده از نرم‌افزار پرکاربرد R در قالب یک کتاب بررسی کنیم. در این کتاب ابتدا نرم‌افزار آماری R به صورت کلی و مقدماتی و با تکیه بر داده‌های ژنتیکی بررسی شده است. انواع داده‌ها و کار با توالی‌های ژنتیکی به عنوان مهم‌ترین بخش آنالیزهای ژنتیکی، توابع ریاضی و آماری مرتبط با ژنتیک، آزمون‌های آماری و استنباطی، انواع نمودارهای پایه و کاربردی در رشته ژنتیکی مورد بحث قرار گرفته و برآورد اثرات ثابت و پیش‌بینی اثرات تصادفی با روش‌های مختلف مطرح شده‌اند. آنالیزهای ارتباط ژنی، پوشش کل ژنوم و پیش‌بینی ژنومی با مثال‌های متعدد تشریح شده‌اند و در پایان مهم‌ترین بخش از آنالیزهای نوین ژنتیکی و ژنومی در سال‌های اخیر به عنوان آنالیز داده‌های نسل سوم بیان ژن به صورت مفصل و در قالب چند فصل به طور کامل و جامع بررسی شده است و انواع کتابخانه‌های مهم و کاربردی در این مباحث به صورت عملی بررسی شده‌اند. در پایان، از همه عزیزان تقاضا داریم ما را در جهت بهبود هرچه بهتر این کتاب یاری کنند.