产品信息

NGS | 数据分析

Dako

Agilent Pathology Solutions

新一代测序数据分析：专为病理学实验室量身打造

Agilent Technologies

前言

新一代测序 (NGS) 已经成为生物与医学研究的常见通用工具。NGS 的成本飞速下降而应用范围愈发广泛，它正在逐步替代许多其他技术。NGS 的高分辨率、低偏差及出色检测能力能够获得先前技术所无法企及的科学发现。利用 NGS 鉴定肿瘤基因组的完整 DNA 序列有望为研究人员理解癌症的起源与演化提供重大突破。

尽管 NGS 应用的范围在过去十年里得到了拓展，但 NGS 数据分析仍然是阻碍其发展成为常规技术的一个主要瓶颈。本文介绍了分析 NGS 数据的基本步骤，包括质量检查和参考基因组映射。我们意识到数据分析和储存方面仍存在巨大挑战，并在试图解决这一难题。我们还阐释了最常见 NGS 应用（变异检测）的进一步数据分析。

普通的 NGS 数据分析软件套装由一级、二级和三级分析工具组成。这种分析包括图像采集、质量控制、碱基识别、与参考基因组的比对、变异识别和生物学解析工具。我们还列出了针对癌症的注意事项。

一级数据分析

在一级数据分析中，原始数据被转化为序列数据（图 1）。在合成测序过程中，碱基对通常在通过激光激发和荧光检测后得到鉴定，并生成相应图像。其他技术直接将化学编码信息（A、C、G、T）翻译成半导体芯片上的数字信息（0、1）。一级分析过程主要用于测序平台，并与测序仪器高度集成。它通常安装在支持测序仪器的本地硬件系统中 (1)。将测序仪器产生的原始信号转换为核苷酸碱基，并最终生成核苷酸序列或“读出序列”。

一级分析还可为每个碱基提供质量值以便后续阶段的分析，类似于 Sanger 测序中使用的 Phred 质量值。Phred 最初开发用于通过 Phred 碱基识别辅助 DNA 测序的自动化，该工具是一个利用荧光“示踪”数据鉴定核碱基的计算机程序。这一过程通常会生成 FASTQ 文件，这种文件是 A、C、G、T 和 N（无碱基识别）一系列序列数据以及每个碱基对应 Phred 质量评分的组合。Phred 质量评分与碱基识别错误概率呈对数相关。例如，30 的 Phred 质量评分表明该碱基识别错误的概率为 1/1000。Q30 是大多数测序数据可以接受的质量评分。它表示碱基正确识别的确定性为 99.9%。这种值被视为高质量数据以及测序机构常用的标准值。

在某些情况下，一级分析还包括标记并合并为单一测序运行的多个样品的多重性分解。标签接头也称为样品条形码，常用于多数当前的 NGS 流程中，允许在测序前对样品进行混合。样品条形码扮演着标识符或标签的作用，以确定读出序列来源于哪个样品。因此可以在一次运行中同时检测多个样品。样品条形码通常是在 DNA 测序前加入的特异性短序列。这些条形码会与未知样品 DNA 一同被测序。测序结束后，将读出序列按照条形码分类和重组（多重性分解）。

此时得到的结果可在二级分析流程中进行处理。

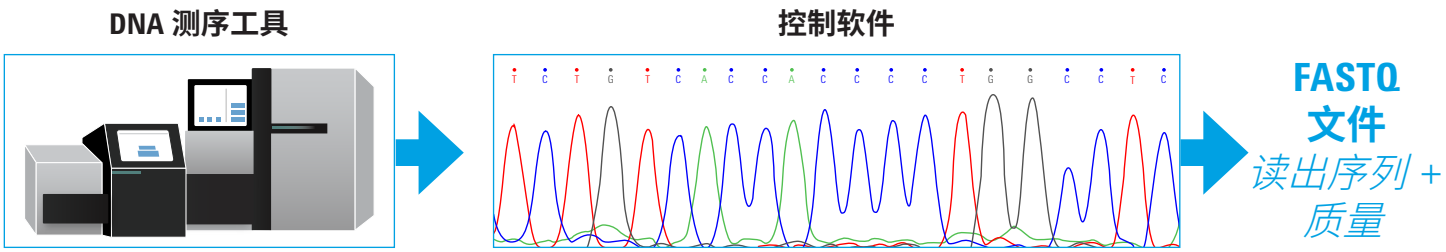

图 1. NGS 数据分析的第一阶段。一级分析：碱基识别

二级数据分析

从 NGS 工具中获得了 FASTQ 格式的原始序列数据后，就会进行读出序列映射或比对的计算密集型步骤（图 2）。将短读出序列映射到参考基因组是比对数据的最标准方式，这一步应尽可能准确，但也需要在合理的时间内完成。由于需要处理的读出序列数量非常多，因此已开发出许多算法来应对这一问题，目前也有很多映射工具。BWA（Burrows-Wheeler 比对）是 DNA 短读出序列最常用的映射工具（表 1）。

将 FASTQ 文件映射到参考基因组后，会生成 SAM 或 BAM 比对文件。SAM 代表序列比对/映射格式。BAM 文件是 SAM 文件的二进制版本，它比 SAM 文件小，通常无法通过人工读取。这些格式已成为报告比对或映射信息的行业标准。

接下来将进行一些精修步骤（如接头切割、软剪切等）。这些步骤中通常包括重复读出序列（可能是 PCR 假阳性结果）的标记/过滤以及重新比对，利用读出序列上的假定插入和缺失综合信息最大程度减少读出序列末端的错误比对。在进入变异识别阶段前，通常在比对数据的基础上对测序软件分配的质量评分进行重新校准。变异识别通过将已测序读出序列与对应参考人基因组上的比对点进行比较，来确定不同统计建模技术（用于辨别真实基因组变异与错误）之间的区域差异 (1)。变异识别的目的是确定样品与参考样品有差异的基因位点。例如，这种变异可能由群体多样性造成，或来自于癌症演化过程中获得的突变。数据通常以 VCF（变异识别格式）文件呈现。如需了解关于 VCF 文件的更多信息，请访问 https://vcftools.github.io/index.html。

与使用上述参考基因组不同，癌症基因组分析中的另一个重要注意事项是需要检测独特的重排，并准确映射个体癌症样品中的染色体断点 (2)。尽管目前用于*从头*组装的算法相当缓慢，但可以进行癌症基因组*从头*组装的软件仍可能成为更强大的工具。

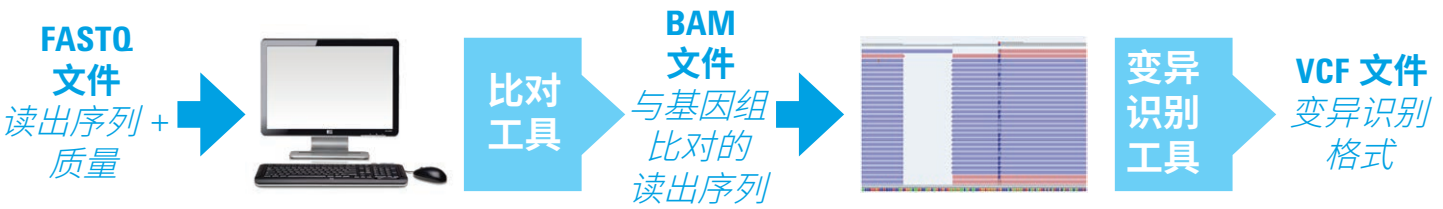

图 2. NGS 数据分析的第二阶段。二级分析：读出序列比对和变异识别

NGS 数据分析中针对癌症的注意事项

处理 NGS 数据的多数生物信息学工具均为正常（即二倍体）基因组而设计，支撑其开发其的假设通常对癌症样品无效，因为癌症样品的 DNA、配对正常比较、肿瘤内克隆异质性和多倍体状态的质量差且数量有限，还因为许多癌症基因组均高度重排 (5)。最近，癌症特异性变异识别程序经过开发后已能够处理 NGS 数据。癌症特异性识别工具包括 Jointsnvmix、Somatic Sniper、MuTect 和 Varscan2（表 1）。安捷伦独立开发的算法称为 SNPPET，这种算法可弥补 SAMtools 及其他无法处理癌症样品的算法的不足。例如，SNPPET 能够可靠地检测低频变异，并完美处理嵌合体样品。

与 SNP（单核苷酸多态性）相比，插入和缺失突变的鉴定更加困难。一些程序（例如 BreakDancer、Dindel 和 Pindel）能够鉴定插入和缺失。由几十到几千个碱基组成的更大型插入和缺失在癌症中十分常见，这需要用特殊的方法进行鉴定。使用配对末端读出序列可从片段的一端或两端进行测序，然后再相互匹配，这是一种解决方案。这种技术还可鉴定融合基因、倒置和易位。

拷贝数变异 (CNV) 和杂合性缺失 (LOH) 对肿瘤形成的作用众所周知。目前还没有公认的工具可以进行这类分析。分析可采取两种方法，将这两种方法相结合可提高准确度。第一种方法将 CNV 看作非常大的插入和缺失，并寻找单个读出序列中的断点、配对读出序列的未对齐或映射过程的整体问题。第二种方法是使用任意位点的读出序列数量作为拷贝数的指标，类似于微阵列技术。

表 1. 用作癌症 NGS 数据序列分析工具的实用免费软件

名称	URL	注释 (5)
映射软件程序		
BWA	http://bio-bwa.sourceforge.net	Burrows-Wheeler 比对工具。
Stampy	http://www.well.ox.ac.uk/stampy	采用 Illumina 读出序列将短读出序列映射到参考基因组。尤其适用于插入和缺失。可与 BWA 结合使用。
Bowtie	http://bowtie-bio.sourceforge.net/index.shtml	Bowtie 是一款存储高效的超快速短读出序列比对工具。
Bowtia2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	Bowtie 2 是一款存储高效的超快速工具，可将读出序列与长参考序列进行比对。
变异识别工具		
GATK	http://www.broadinstitute.org/gsa/wiki/	可作为变异识别的金标准。具有分析 NGS 数据程序的结构化软件库。可用于变异识别和插入缺失鉴定。
JointSNVMix	http://code.google.com/p/joint-snv-mix/	可同时分析肿瘤与正常基因组配对，可鉴别生殖系与体细胞突变。
MuTect	http://www.broadinstitute.org/cancer/cga/mutect	一款可从肿瘤正常配对测序数据中鉴定体细胞点突变的变异识别工具。这款程序可由肿瘤和正常样品的覆盖深度确定是否有足够灵敏度来识别体细胞突变。
SAMtools	http://samtools.sourceforge.net/	用于操作已比对数据（包括 SNP 查找）的工具。
Somatic Sniper	http://gmt.genome.wustl.edu/somatic-sniper/current/	这款程序通过比较肿瘤与正常数据生成基于 Phred 的概率评分，以确定肿瘤和正常基因型存在差异的可能性。
Varscan2	http://dkoboldt.github.io/varscan/	可用于鉴定体细胞和生殖系变异以及肿瘤正常配对中的 LOH 事件。已用于鉴定肿瘤正常外显子数据中的 CNV。这是一款不受平台影响的工具，可用于包括 Ion Torrent 在内的多数 NGS 平台的数据处理。
Freebayes	https://arxiv.org/abs/1207.3907	贝叶斯基因变异检测器，设计用于查找小型多态性，尤其是 SNP、插入和缺失、MNP（多核苷酸多态性）以及小于短读出序列测序比对长度的复杂事件（插入和替换复合事件）。
SNPPET	http://www.agilent.com/genomics/surecall	包含在安捷伦数据分析软件 SureCall 内。请参阅 SureCall* 手册了解 SNPPET 的详细描述。
插入缺失和结构变异识别工具		
BreakDancer	http://breakdancer.sourceforge.net	BreakDancer Max 可通过标注配对末端读出序列鉴定结构变异，这些读出序列的映射距离无法预测或方向错误。检测大型插入、删除、倒置、染色体内/间易位。BreakDancer Mini 可用于检测 10-100 bp 的小型插入缺失。
Dindel	https://sites.google.com/site/keesalbers/soft/dindel	发现小型插入缺失。在覆盖更深入的情况下，通过过滤数据降低假阳性数量，以确保每个插入缺失出现两次以上。
Genome STRiP	http://www.broadinstitute.org/software/genomestrip/genome-strip	设计用于检测多个体共有的结构变异。获得满意结果需要 20-30 个基因组。其当前用途仅限于对相对于参考序列的删除片段进行发现和基因分型。
Pindel	http://gmt.genome.wustl.edu/packages/pindel/	可用于鉴定简单的删除和插入。利用配对末端读出序列来鉴定大型断点和中型插入。可检测倒置和串联重复。

三级数据分析

三级数据分析（即解析）是 NGS 数据分析流程中最复杂、最耗时的实验特异性手动阶段。具体来看，生殖系全外显子组测序 (WES) 平均产生约 20000 个 SNP (3)，Agilent SureSelect^{XT2} Human 全外显子 V6+COSMIC * 甚至会产生 30000 个 SNP，这些都需要进行过滤。这些 SNP 大多是同义或良性变化。其余的罕见变异有待进一步研究，可分为致病性变异、良性变异或 VUS（“未知临床意义的变异”）。靶向基因组合中发现的变异数量要低得多，但仍需要进行过滤。许多用于常见变异标注的数据库均已公开发布，如 1000 Genomes Project (http://www.1000genomes.org/) 以及 dbSNP 数据库 (http://www.ncbi.nlm.nih.gov/projects/SNP/)。

用于变异解析的多种工具和数据库均已上市销售。还有一些预测变异功能的软件，如 SIFT 和 PolyPhen。表 2 中列出了 CIViC 等可用于变异标注和解析的工具和资源。

Agilent Cartagenia Bench Lab for Molecular Pathology 包含许多此类工具和数据库，可以让您有效优化体细胞变异并过滤相关候选变异，以在临床环境下进行回顾和评估，其中包括生物医学相关的先前研究结果和信息。

表 2. 预测突变蛋白功能和变异解析的资源

程序	描述	URL
Polyphen-2	突变功能预测	http://genetics.bwh.harvard.edu/pph2
SIFT	突变功能预测	http://sift.jcvi.org
CHASM	突变功能预测	http://wiki.chasmsoftware.org
ANNOVAR	标注	http://www.openbioinformatics.org/annovar/
COSMIC	癌症体细胞突变目录	http://www.sanger.ac.uk/genetics/CGP/cosmic
UCSC 癌症基因组学浏览器	基于网页的工具，用于可视化、整合和分析癌症基因组学及相关临床数据	https://xenabrowser.net/heatmap/
Cancer Genome Workbench	来自 TCGA、TARGET、COSMIC、GSK 和 NCI60 计划的宿主突变、拷贝数、表达和甲基化数据；用于不同癌症中样品级基因组学和转录变化的可视化工具	https://cgwb.nci.nih.gov/
HGVS	人类基因组变异协会；变异标注的推荐工具	http://varnomen.hgvs.org/
RefSeq 数据库	编码序列的衍生	www.ncbi.nlm.nih.gov/RefSeq
dbSNP	单核苷酸多态性数据库	http://www.ncbi.nlm.nih.gov/projects/SNP
HGMD	人类基因突变数据库	http://www.biobase-international.com/product/hgmd
ClinVar	整合了关于基因组变异及其与人类健康关联的信息	https://www.ncbi.nlm.nih.gov/clinvar/
CIViC	癌症变异的临床解析	https://civic.genome.wustl.edu/#/home

数据可视化

一般来说，不需要查看原始数据。虽然大部分分析可自动完成，但人工解析、科研经验和判断有时仍十分有用。例如，单核苷酸变异识别相对较为稳定。但插入缺失或基因融合可能会出现问题：一些插入缺失读出序列由于无法与参考序列进行适当对齐而被舍弃。

综合基因组学查看器 (IGV) 是一款高性能浏览器，可高效处理大型数据集 (4)。例如，BAM 文件、VCF 文件和 BED 文件均可在 IGV 中显示。IGV 是一款以 Java 编程语言编写的桌面应用程序，可在所有主要平台（Windows、Mac 和 Linux）上运行。它的主要任务是为希望对自己或同事的数据集进行查看和探索的研究人员提供支持。具有 GNU LGPL 开源许可证时，可以在 http://www.broadinstitute.org/igv 免费下载 IGV。IGV 窗口可划分为图 3 所示的多个控件和面板。顶部带控件的命令栏可用于选择参考基因组、导航以及定义目标区域。命令栏下方的标题面板可展示当前浏览染色体的表意符号以及表示显示区域大小的基因组坐标尺。窗口的其余部分被划分成一个或多个数据面板和属性面板。

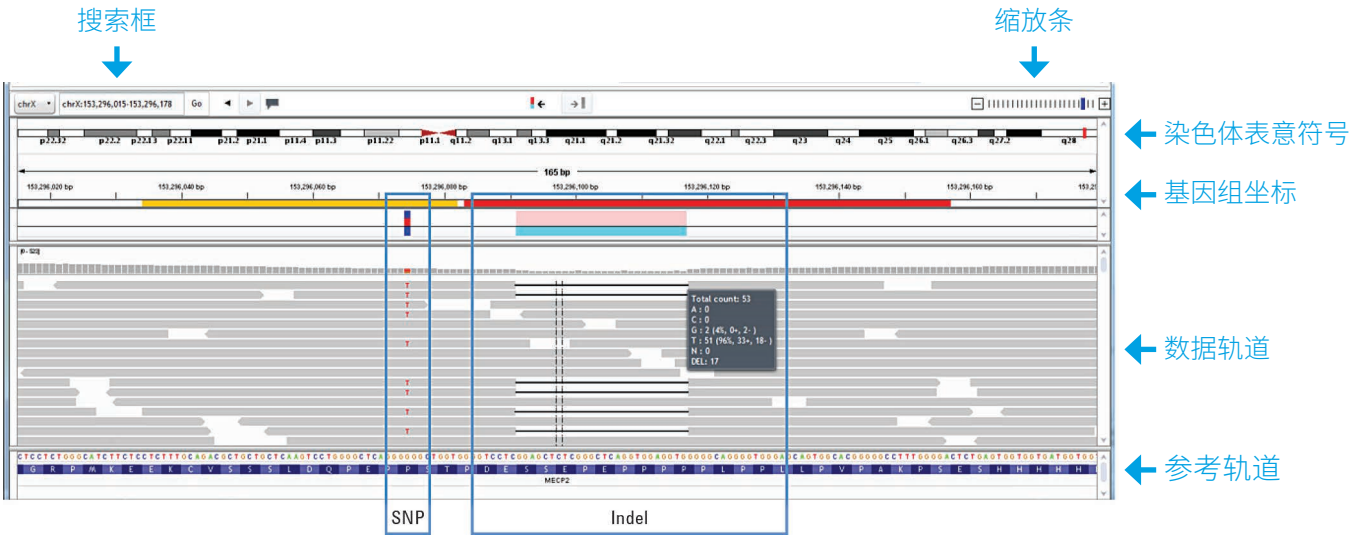

图 3. IGV 应用程序窗口

IT 基础架构

所需的 IT 基础架构很大程度上取决于测序数据集的大小和体积 (6)。某些数据集可能十分巨大。例如包括原始序列、比对和变异识别的人类全基因组测序项目，其中每个样品的数据量可以达到数百 GB。然而，样品的靶向测序数据集通常要小得多，可能只有几 MB 或几 GB 大小。

除存储空间外，另一个关键资源是计算能力。这些数据集过于庞大而无法使用台式电脑对其进行有效分析，尤其是比对。普通用户通常需要 16 到 24 GB 的内存来比对序列数据。

许多大学或研究机构提供了集群资源。一个集群可看作是大量小型计算机并行联网形成的一种专用超级计算机。

对于无权访问集群的研究人员来说，可以选择一些基于网络和云端的替代方案。其中一个基于云端的替代方案的例子就是亚马逊网络服务 (AWS)。这样的系统让云服务提供商进行繁重的计算密集型操作，从而提供了高度灵活性。此外，基于网络或云端的解决方案令购买、维护和升级 IT 基础架构不再困难。对于更青睐端到端解决方案的用户来说，Agilent Cartagenia Bench Lab for Molecular Pathology 是一种理想的商业化选择。

如需了解更多信息和资源，请访问

NGS 癌症资源中心
http://www.genomics.agilent.com/article.jsp?pageId=8200008&CID=G012110

Cartagenia Bench Lab for Molecular Pathology
http://www.agilent.com/en-us/solutions/clinical-grade-variant-assessment/cartagenia-bench-lab-for-molecular-pathology

文件格式术语表

名称	描述
FASTA	FASTA 文件格式是一种简单的文本格式，其中包括含由起始指示符（通常是“>”）、注释和字母编码的核苷酸序列组成的一条或多条记录。
FASTQ	FASTQ 文件格式最初由韦尔科姆基金会桑格学院研究所开发用于处理 FASTA 序列及其质量数据。
SAM	SAM 是一系列序列及其与参考基因组的比对。SAM（序列比对/映射）格式是一种非常通用且近乎标准化的格式，用于储存多个已比对核苷酸序列。
BAM	BAM 也是一系列序列及其与参考基因组的比对，但它是 SAM 格式更紧凑的二进制形式。
VCF	经过 1000 Genomes Project 标准化的样品和参考基因组之间的变异。
BED	BED 文件 (.bed) 是制表符分隔的文本文件，它可以确定特征轨迹。

参考文献

(1) Oliver *et al.* Bioinformatics for Clinical Next Generation Sequencing. *Clinical Chemistry* (2015) 61: 124-135.

(2) Gullapalli *et al.* Next Generation Sequencing in Clinical Medicine: Challenges and Lessons for Pathology and Biomedical Informatics. *J Pathol Inform* (2012) 3:40.

(3) Chang *et al.* Clinical Application of Amplicon-Based Next-Generation Sequencing in Cancer. *Cancer Genetics* (2013) 206: 413-419.

(4) Thorvaldsdottir *et al.* Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration. *Briefings in bioinformatics* (2012) 14: 178-192.

(5) Ulahannan *et al.* Technical and Implementation Issues in Using Next-Generation Sequencing of Cancers in Clinical Practice. *British Journal of Cancer* (2013) 109: 827-835.

(6) Perkel *et al.* Sequence Analysis 101. *The Scientist* (2011) March 1.

可靠结果，完整方案。

Dako

Agilent Pathology Solutions

www.agilent.com

我们的客户代表遍布 100 多个国家/地区

澳大利亚
+61 1800 802 402

奥地利
+43 1 408 43 34 0

比利时
+32 16/93 00 30

巴西
+55 11 50708300

加拿大
+1 800 387 8257

中国
+80 08 20 3278

丹麦
+45 44 85 97 56

芬兰
+358 9 348 73 950

法国
+33 1 64 53 61 44

德国
+49 40 69 69 470

爱尔兰
+353 1 479 0568

意大利
+39 02 58 078 1

日本
+81 3 5232 9970

韩国
+82 80 004 5090

新西兰
+31 20 42 11 100

挪威
+47 23 14 05 40

波兰
+48 58 661 1879

西班牙
+34 93 344 57 77

瑞典
+46 8 556 20 600

瑞士
+41 41 760 11 66

英国
+44 (0)1 353 66 99 11

美国
+1 805 566 6655

29169 2017JAN10

PR7000-0612

** 仅限研究使用。不可用于诊断目的*

2017 年 1 月 13 日，中国出版

5991-7805CHCN

Agilent Technologies