# 测序技术的个体化医学检测

# 应用技术指南（试行）

# 前　言

随着人们对疾病分子机理的认识，以及测序技术的不断发展和完善，已促使基因测序技术走入临床实验室。通过基因测序的方法可以对较长的基因片段进行检测，并可一次性发现基因多态性、点突变、小片段缺失等多种复杂基因变异类型，因此已经被广泛应用多种分子诊断领域，包括癌症检测（遗传性癌症的致病基因检测、癌症易感性基因、或靶向性抗肿瘤药物的靶点基因检测等），遗传学检测（遗传病的诊断和携带者检测），药物基因组学检测（根据药代动力学和/或药效动力学相关基因的遗传背景差异，为患者量体裁衣式地提供疗效更好、毒副作用更低的个体化治疗方案）和微生物检测（病毒基因分型，耐药基因检测）等。

近年来随着新一代测序（next generation sequencing，NGS）技术的不断发展和检测成本的持续下降，NGS 也已经发展成为重要的临床基因分析技术。NGS 的本质是大规模平行测序（massively parallel sequencing，MPS），作为一个广义的范畴，包括许多原理各异、但都可以一次性产生大量数字化基因序列的多种测序技术。相对 Sanger 测序而言，NGS 技术流程较为复杂，最重要的区别就是必须依赖下游的生物信息学处理过程，才能将从测序仪获得的原始数据转换成可以用于临床解读的 DNA 序列，因此需要同时配备专业的技术人才、临床专家和信息分析专家，而且对场地和环境要求高。目前很多大中型医学实验室已经开始使用 NGS 开发各种检测项目，包括遗传疾病诊断，癌症和传染性疾病基因检测等。其他还有可能用于临床的检测项目还包括全基因组甲基化分型、微生物组、宏基因组/泛基因组，以及转录组测序等。

目前已经用于临床检测的基因测序项目绝大多数都属于实验室自主研发项目 （laboratory developed tests，LDT），因此，为了确保所获得的序列和结果分析能够用于指导临床决策，需要对测序的全过程，包括样品处理、检测步骤和数据解读等诸多方面进行标准化。本指南从样品采集、运输、接收、处理、检测、检测项目的开发、检测的验证（verification）与确认（validation）、室内及室间质量控制需遵循的基本原则、结果报告和解释，以及可能出现的问题及应对措施等，为基于测序技术的个体化医学检测应用提供标准化指导。

# 目　录

## 1. 摘要

采用测序技术进行基因分型已经从实验室研究进入了临床应用和个体化医学指导。测序是对复杂基因变异类型进行基因分型的首选检测技术，特别是对当数以百计或数千的较长基因序列进行检测分析更有优势。本指南针对使用自动毛细管电泳 Sanger 测序技术和新一代测序技术的测序项目，制定了样品采集和处理测序过程、序列比对和组装、检测确认和验证、持续的质量保证以及结果报告等各项标准。

**关键词**

毛细管电泳；Sanger 测序；大规模平行测序；新一代测序；聚合酶链反应；基因检测，质量控制。

## 2. 本指南适用范围

本指南由国家卫生计生委个体化医学检测技术专家委员会制定，是国家卫生计生委个体化医学检测指南的重要内容，旨在为基于测序技术的个体化医学检测应用提供标准化指导。本指南的主要适用对象为所有采用基因测序的方法、开展个体化医学分子检测的临床检验实验室。

鉴于测序技术一直在不断升级和更新，因此本指南的技术标准尽可能针对所有的测序平台，所讨论和描述的问题不仅限于某一个特定的测序平台。

本指南对于测序技术和序列结果的解读提供通用指南，但不包含针对某一项特定的临床应用项目提供具体的测序技术和数据分析标准以及结果解读意见。

本指南适用于胚系突变（germline mutation）、体细胞突变（somatic mutation）以及微生物基因组相关的个体化医学检测。

## 3. 简介

测序技术是用于分子诊断中基因分型的重要手段，目前常规应用于遗传病的诊断与携带者筛查、感染性疾病病原微生物（例如 HIV、HCV 等）的基因分型、器官移植时的 HLA 高分辨率分型、肿瘤的遗传背景筛查及个体化用药基因检测等。这些基因测序项目绝大多数都属于实验室自主研发项目，因此需要对样

品处理、检测步骤、质量保证和质量控制、以及数据解读等测序全过程进行标准化，才能确保所获得的序列和结果分析能够用于指导临床决策。

除了经典的基于双脱氧终止和毛细管电泳技术的 Sanger 测序技术，近年来新一代测序（NGS），又称大规模平行测序（MPS），已经发展成为重要的临床基因分析技术。目前基于 NGS 的临床应用也越来越多，例如针对遗传性疾病的多基因检测组合（multigene panels）、针对一些具有重要临床意义的低含量突变（例如 HIV 耐药基因突变）、特别是针对肿瘤的 NGS 检测，覆盖多个肿瘤相关基因、检测低含量的突变，对于肿瘤的精准和分层诊断、靶向药物的选择等都具有重要临床指导意义。当测序价格更低、数据分析更全面后，表观遗传学、微生物基因组、宏基因组学等都将在临床应用中发挥重要作用。鉴于 NGS 技术流程的复杂性，在测序平台确认、测序过程、结果确认、数据分析和管理、以及质量控制参数等方面也提出了更高的要求。

从测序仪获得的原始数据可以转换成用于临床解读的 DNA 序列，这就需要同时配备专业的技术人才、临床专家和信息分析专家，而且对场地和环境要求高。目前很多大中型医学实验室已经开始使用 NGS 开发各种检测项目，包括遗传疾病诊断，癌症和传染性疾病基因检测等。其他还有可能用于临床的检测项目还包括全基因组甲基化分型、微生物组、宏基因组/泛基因组，以及转录组测序等。

## 4. 标准声明/警告

由于难以判断哪些类型的样品具有传染性，所有受检对象和临检样品都应视为具有传染性。实验室的生物安全性非常重要，临检操作中样品的采集和处理过程中应遵循 GB 19781-2005《医学实验室安全要求》、GB 19489-2008《实验室生物安全通用要求》和卫生部《微生物和生物医学实验室生物安全通用准则》，确保实验工作人员的安全和实验活动的顺利进行，避免对检测工作人员和环境可能造成的危害。

## 5. 标准术语和缩略语

### 5.1 标准术语

#### 5.1.1 基因

是遗传物质的最小功能单位，是指具有一定生物学意义的一段 DNA。

### 5.1.2 脱氧核糖核酸 DNA

核酸的一种，是由特殊序列的脱氧核糖核苷酸单元（dNTP）构成的多聚核苷酸，起携带遗传信息的功能。DNA 为一种双链分子，通过核苷酸碱基对间较弱的氢键维系。DNA 包含的 4 中核苷酸包括：腺嘌呤（A）、鸟嘌呤（G）、胸腺嘧啶（T）和胞嘧啶（G）。人类存在两种类型的 DNA：来自细胞何种染色体的基因组 DNA（gDNA）和线粒体 DNA。

### 5.1.3 核糖核酸 RNA

核糖核酸，与 DNA 类似的单链核酸，由核糖核苷酸按照一定的顺序排列而成，含尿嘧啶而不含胸腺嘧啶，存在于细胞质和细胞核中，在细胞的蛋白质的合成和其他化学活动中起重要的作用。RNA 分子包含信使 RNA（mRNA）、转运 RNA（tRNA）、核糖体 RNA（rRNA）和其他小 RNA 等多种类型，分别行使不同的功能。各种 RNA 的混合物称为总 RNA。

### 5.1.4 基因型 genotype

又称遗传型，是某一生物个体全部基因组合的总称，它反映生物体的遗传构成，即从双亲获得的全部基因的总和。据估计，人类的结构基因约 5 万对。因此，整个生物的基因型是无法表示的，遗传学中具体使用的基因型，往往是指某一性状的基因型。

### 5.1.5 等位基因 allele

一般是指位于一对同源染色体相同位置上控制某一性状的不同形态的一对基因。若成对的等位基因中两个成员完全相同，则该个体对此性状来说成为纯合子。若两个等位基因各不相同，则该个体对该性状来说是杂合子。

### 5.1.6 基因组合 gene panel

是指可以导致共同临床表型的一组基因或变异体，可以采用靶向区域捕获和新一代测序技术同时进行检测。

### 5.1.7 变异 variation

是指 DNA 序列中与参考序列不同的任何核苷酸序列改变。

### 5.1.8 单核苷酸多态性 SNP

是指 DNA 序列中单个核苷酸—A，T，C 或 G—的变异，在人群中的变异频率>1%，造成包括人类在内的物种之间染色体基因组的多样性，通常不会导致严重的临床表型。

### 5.1.9 单倍型 haplotype

是单倍体基因型的简称，在遗传学上是指在同一染色体上进行共同遗传的多个基因座上等位基因的组合，由若干个决定同一性状、紧密连锁、具有统计学关联性的单核苷酸多态性构成。

### 5.1.10 插入/缺失 insertion/deletion（indel）

是指与参考序列相比，存在一定数量的核苷酸插入或缺失，也可以在同一部位同时发生缺失和插入突变。

### 5.1.11 拷贝数变异 copy number variant （CNV）

是指 1 kb 及以上的 DNA 大片段的插入或缺失。

### 5.1.12 偶发突变 incidental findings

在新一代测序检测中发现的一些与患者预期的临床表型无关联的基因突变。

### 5.1.13 光密度

表示被检测物吸收的光密度，260nm 下的吸光值可用来表示 DNA 的相对浓度，具体换算是 DNA 浓度（ng/μl）=$OD_{260}$×50×稀释倍数。

### 5.1.14 富集 enrichment

通过特定的方法使得混合细胞或核酸样品中待测核酸的比例增加，例如可以将不想检测的核酸成分选择性去除、或者对待测核酸成分进行选择性探针捕获或 PCR 扩增。

**5.1.15 杂交捕获 capture by hybridization**

通过固相表面的互补 DNA 序列与靶向 DNA 进行杂交，对待测核酸进行靶向富集的方法。

**5.1.16 基因组文库 genomic library**

将某生物的全部基因组DNA切割成一定长度的DNA片段克隆形成的集合，可用于下游的扩增、Sanger 测序或新一代测。

**5.1.17 聚合酶链反应 polymerase chain reaction（PCR）**

是一项对特定的 DNA 片段进行体外酶促快速扩增的分子生物学技术。通过模拟 DNA 的自然复制过程，引物按照碱基配对与 DNA 模板互补结合以后，在 DNA 多聚酶的作用下，按照碱基配对的原则（A 对 T，C 对 G），从引物开始合成与模板 DNA 互补的 DNA 链。

**5.1.18 乳液 PCR emulsion polymerase chain reaction（ePCR）**

是将单分子 DNA 片段分配到一个乳滴进行 PCR 扩增，可以同时保证多重 PCR 扩增产物的丰度和纯度。

**5.1.19 测序 sequencing**

分析特定 DNA 片段的碱基序列，也就是腺嘌呤（A）、胸腺嘧啶（T）、胞嘧啶（C）与鸟嘌呤的（G）排列方式。

**5.1.20 新一代测序 next generation sequencing（NGS）**

又称大规模平行测序 massively parallel sequencing（MPS），是指采用“边合成边测序”的原理、对于几十万到几百万 DNA 分子同时进行平行的测序反应，然后通过生物信息学分析所得到的原始图像数据或电化学信号、最终得到待测样品的核酸序列或拷贝数等信息的测序技术，又称为高通量测序、深度测序等。

**5.1.21 靶向区域测序 target sequencing**

对于特定的核酸区域，例如特定基因组合、外显子等，进行检测和测序。

**5.1.22 外显子组 exome**

至某个体的全部编码区域外显子。例如全外显子测序（whole exome sequencing，WES）就是指对基因组中的全部蛋白质编码区的序列进行测序。

**5.1.23 宏基因组学 metagenomics**

通过直接从特定微环境样品中提取全部微生物的 DNA，不需要事先对每一种菌群进行单独分离，直接构建宏基因组文库，分析其中所包含的全部微生物的遗传组成及其群落功能。是在微生物基因组学的基础上发展起来的一种研究微生物多样性的新理念和新方法。

**5.1.24 微生物组 microbiome**

指存在于人体特定区域（如皮肤、肠道等）的全部微生物群体。

**5.1.25 映射 mapping**

是指通过将测序片段定位以产生基因图谱的过程。

**5.1.26 比对 alignment**

是指根据两个或多个的核苷酸序列的重合部分，来构建连续的核酸序列，或者据此找出序列结构变化的错配、插入、缺失和易位部分。

**5.1.27 重叠群 contig**

是指在基因组测序过程中，将许多短的序列片段交叠链接而成的连续的、不间断的 DNA 片段。

**5.1.28 原始数据 raw data**

未经处理的原始测序数据。

**5.1.29 碱基识别 base calling**

根据得到的原始数据，运用计算机读出软件（base calling software） 进行图象或电化学信号处理，以确定原始数据所蕴含的 DNA 序列。

**5.1.30 有效数据 clean data**

去除了接头和低质量序列的数据。

**5.1.31 覆盖深度 coverage/depth of coverage**

用于特定区域碱基识别的有效核酸测序片段，又称读段（reads）的数目。

**5.1.32 扩增偏倚 amplification bias**

是指在对模板 DNA 进行 PCR 扩增过程中，因为扩增效率的不同，某些部分比其他区域会生成更大量的拷贝的现象。在 NGS 过程的样品处理过程中，扩增偏倚可能导致后续的测序结果质量下降和解读错误。.

**5.1.33 PCR 始祖效应 PCR founder effect**

在初始 PCR 扩增循环中得到扩增的特定 DNA 分子在后期测序读段中占有优势比例的现象。

**5.1.34 质量值 Quality Score（Q-score）**

测序时碱基识别（Base Calling）过程中，对所识别的碱基给出的错误概率。

**5.1.35 条码 barcode**

是指一段特征性的脱氧核苷酸短片段；在多样品混合检测时，充当一个识别特定样品来源的唯一标志；也有测序平台将其称为标签（index）。

**5.1.36 标签互换 index swapping**

因为测序数据读取错误所致的条码关联错误。

**5.1.37 接头 adapter/ oligonucleotide adapter**

用于偶联寡核苷酸片断的脱氧核苷酸短片段。

**5.1.38 检出限 limit of detection（LOD）**

样品中一种分析物可被检出的最低的含量，这一分析物含量有可能不是量化的具体数值。

**5.1.39 线性**

在已知的范围内，某检测提供的结果能够直接与样品的浓度（或量值）成比例关系的能力。

**5.1.40 重复性 repeatability**

是用本方法在正常和正确操作情况下，由同一操作人员，在同一实验室内，使用同一仪器，并在短期内，对相同试样所作多个单次测试结果，在 95%概率水平两个独立测试结果的最大差值。

**5.1.41 重现性 reproducibility**

是指在不同实验室由不同分析人员测定结果之间的精密度。

**5.1.42 准确度 accuracy**

在一定实验条件下测定的结果与真值相符合的程度，即对于核酸序列的分析结果与参考序列的一致性。主要受系统误差的影响。

**5.1.43 精密度 precision**

对于给定样品进行重复检测时，检测结果在同一批次内的可重复性（repeatability）和不同批次之间的可重复性（reproducibility）。

**5.1.44 灵敏度 sensitivity**

检测系统或仪器对被测物的变化所发生的相应变化，即对于已知核酸序列异常的检出能力。

**5.1.45 特异性 specificity**

能专一检测被测物的试验能力，即当序列中不存在的核酸异常不应该被检测出来，实际验证中将采用假阳性率进行考核。

**5.1.46 报告范围 reportable range**

是指具有可以接受的质量水平的基因组区域，常用于指测序技术所检测的的 DNA 区域或其他特定的核酸变异形式。

**5.1.47 参考范围 reference range**

测序技术所检测的核酸区域中可能出现在正常人群中的序列变异。

**5.1.48 验证 verification**

通过提供客观证据对规定要求已得到满足的认定。

**5.1.49 确认 validation**

通过提供客观证据对特定的预期用途或应用要求已得到满足的认定。

**5.1.50 证实检测 confirmatory testing**

通过另外一种特异性和/或敏感性更强的检测方法，对于检测结果进行验证。

**5.1.51 临床相关性 clinical relevance**

是指基因组学改变或特征对于患者的诊断或管理具有指导意义。

**5.1.52 临床有效性 clinical validity**

是指检测与感兴趣的临床特征之间的关联强度，通常用临床敏感性和临床特异性来表示。对于基因检测，临床有效性就是指基因型和表型之间的相关性。.

**5.1.53 室内质控**

实验室内进行的用于满足质量要求的操作技术和活动。

**5.1.54 室间质量评价**

通过实验室间比较来评价实验室的检测能力，又叫做能力验证。

**5.1.55 能力验证 proficiency testing（PT）**

通过实验室间比较来评价实验室的检测能力，同室间质量评价的定义。

**5.1.56 质控品**

指专门用于质量控制目的的样品或溶液。

**5.1.57 知情同意**

患者有权利知晓自己的病情，并可以对医务人员所采取的治疗措施和临床检测项目决定取舍的权利。

**5.1.58 实验室自主研发的检测项目 laboratory developed tests（LDT）**

是指仅在实验室内部使用、不外售给其他实验室或医疗机构、但结果可以

用于指导临床诊疗的检测项目。根据美国的管理体系，经过 CLIA（《临床实验室改进修正案》）认证的实验室即获得 LDT 经营许可权限。

**5.2 缩略语**

A：adenine　腺嘌呤

bp：base pair(s) 碱基对

C：cytosine　胞嘧啶

CDS：coding DNA sequence　编码 DNA 序列

CFDA：china food and drug administration 国家食品药品监督管理总局

CNV：copy number variant 拷贝数变异

CRT：cyclic reversible termination 循环可逆性末端终止

DNA：deoxyribonucleic acid 脱氧核糖核酸

dNTP：deoxy-ribonucleoside triphosphate 脱氧核苷三磷酸

EDTA：ethylenediaminetetraacetic acid　乙二胺四乙酸

EQA：external quality assessment 室间质量评价

FDA：food and drug administration 美国食品药品监督管理局

FFPE：Formalin fixed and paraffin embedded 甲醛固定与石蜡包埋

G：guanine 鸟嘌呤

HCV：hepatitis virus C　丙型肝炎病毒

HIV：Human Immunodeficiency Virus 人类免疫缺陷病毒

HGNC：human gene nomenclature committee 人类基因命名委员会

HGVS：human genome variation society 人类基因组变异学会

HLA：　Human leukocyte antigens　人类白细胞抗原

indel：insertion/deletion 插入/缺失

LDT：laboratory developed test 实验室自主研发项目

LoD：limit of detection 检出限

PBMC：peripheral blood mononuclear cell 外周血有核细胞

PCR：polymerase chain reaction 聚合酶链式反应

PT：proficiency testing 能力验证

QA：quality assurance 质量保障

QC：quality control 质量控制

RefSNP allele：参考 SNP 等位基因

RNA：ribonucleic acid 核糖核酸

SNP：single nucleotide polymorphism 单核苷酸多态性

SNV：single nucleotide variation 单核苷酸变异

SOP：standard operation procedure 标准操作规程

T：thymine 胸腺嘧啶

Ti/Tv：transition/transversion ratio 转换/颠换率

U：uracil 尿嘧啶

ZMW：zero-mode waveguide 零级波导

## 6. 测序技术概述与应用

### 6.1 测序技术概述

成熟的 DNA 测序始于 20 世纪 70 年代中期的 Sanger 双脱氧链终止法与 Maxam-Gilbert DNA 化学降解法测序，此后，Sanger 测序方法更受欢迎并在后续得到一系列改进，首先是使用四色荧光染料代替放射性核素对 ddNTP 的标记，比传统的放射性同位素方法容易，而且安全、快速、成本低。其次是采用毛细管电泳技术分离 DNA 片段，使测序得以自动化进行，通过缩短运行时间、增加读

长、增加便利性等使得测序的安全性和通量均大幅提高，目前被全世界很多实验室广泛使用。

尽管传统的 Sanger 法测序具有阅读长度长、精确度高等目前仍无法超越的优点，但由于其单次测序的通量较低，因此在检测较大的基因组片段的基因序列时仍然存在成本高、速度慢等缺点，并不是最理想的测序方法。NGS 技术通过高通量的平行测序反应结合后期的生物信息学数据处理，使得测序速度大大提高，而检测成本则大幅下降。但是，NGS 技术复杂、对场地和环境要求高，还需要同时配备专业的技术人才、临床专家和信息分析专家，因此，需要建立更为详尽的质量管理体系和标准，包括系统验证、实验室内部质量控制、外部质量考核与评价以及能力验证等，进行 NGS 临床应用的规范和指导。

目前商业化生产的 NGS 检测平台有多种，技术原理以及优、缺点也各有不同（附录 A 表 1）。

### 6.1.1 双脱氧链终止法测序（Sanger 测序）

1977 年 Sanger 等发明的 DNA 双脱氧链末端终止测序法。其基本原理为：利用双脱氧核苷三磷酸（ddNTP）缺乏延伸所需的 3-OH 基团这一特点，将 ddNTP 作为链终止试剂，通过 DNA 聚合酶的引物延伸产生一系列不同长度的 DNA 片段，再进行分离。测定时，首先将模板分在四个 DNA 反应体系，再分别加入引物，DNA 聚合酶，所有四种脱氧核苷三磷酸（dNTP），并分别混入一定比例的带有放射性核素标记的某种双脱氧核苷三磷酸（ddNTP）。以高分辨率凝胶电泳分离获得一系列大小不同的 DNA 片段后，就可以通过放射自显影确定所测的 DNA 序列。此后，在 Sanger 测序法又经过不断发展和改进，80 年代中期出现了以荧光素标记代替放射性核素标记、以荧光信号接收器和计算机信号分析系统代替放射性自显影的自动测序仪，90 年代中期出现的毛细管电泳技术使得测序的通量大为提高。现今的 Sanger 测序技术也已实现了自动化，采用四色荧光染料代替放射性核素对 ddNTP 的标记，毛细管电泳分离 DNA 片段，使测序的便利性，安全性及获得的通量均大大提高。

Sanger 测序技术在人类基因组计划 DNA 测序的后期阶段起了关键作用，加速了人类基因组计划的完成。经过了 30 年的不断发展与完善，现在已经可以对

长达 1,000bp 的 DNA 片段进行测序，对每一个碱基的读取准确率高达 99.999%，测定每千碱基长度序列的成本是 0.5 美元，每天的数据通量可以达到 600, 000 bp。尽管由于对电泳分离技术的依赖，第一代测序技术在速度和成本方面都已达到了极限，但因其久经考验的准确性和初具规模的市场占有率，Sanger 测序目前仍然是基因测序的金标准。

### 6.1.2 焦磷酸测序

焦磷酸测序技术最早由瑞典皇家科学院的 Nyren Pal 于 1987 年提出，1998 年，Ronaghi 等在《Science》上首次报道了这项技术。它是一种实时定量的 DNA 测序技术，其原理可以概括为“边合成边测序”。其核心原理是测序引物与单链 PCR 产物结合后，4 种原料 dNTP 模板发生碱基配对反应形成共价键，该 dNTP 的焦磷酸基团（PPi）被释放出来，且 PPi 的量与结合的 dNTP 量呈正比。然后，底物 5’-磷酰硫酸（APS）在 ATP 硫化酶的催化下与 PPi 形成等量的 ATP，ATP 又为荧光素酶提供能量，介导荧光素转化成氧化荧光素并发出与 ATP 的量呈正比的可见光信号，这些光信号最终形成峰图，峰高与合成反应中掺入的核苷酸数目呈正比。而 ATP 和未参加反应的 dNTP 由双磷酸酶降解，淬灭光信号后加入下一种 dNTP 继续下一轮的反应。随着循环反应的进行，我们便可通过信号峰的有无判断碱基的种类，通过信号峰的峰高检测碱基的数目等 DNA 序列信息。从开始提出至今，焦磷酸测序技术不断优化，逐步发展成为一种高通量、高精确度、高稳定性的实时测序技术，后来 Roche 公司 454 技术使用的测序方法的原理就是焦磷酸测序技术。

### 6.1.3 新一代测序（next generation sequencing，NGS）

NGS 又称大规模平行测序（MPS），包含多种可以一次性产生大量数字化基因序列的测序技术，是继 Sanger 测序的革命性进步，采用平行测序的理念，能够同时对上百万甚至数十亿个 DNA 片段进行测序，实现了大规模、高通量测序的目标。NGS 由模板制备和序列检测过程（湿实验部分）和数据分析过程（干实验部分）两部分组成。由 NGS 所产生的大量测序数据需要复杂的生物信息学工具进行分析和解读，以及用于存储和管理的计算机资源。因为 NGS 在速度、通量和价格方面均具有明显的优势，而且可以同时对多个基因区域的基因变异进

行识别、灵敏检测低含量的突变，已使得 NGS 技术在分子诊断、医药健康等领域展示出广阔的应用前景。

目前商业化生产的 NGS 平台有多种，技术原理以及优、缺点也各有不同（附录 A 表 1），主流的 NGS 技术主要有基于焦磷酸测序原理的 454 测序技术、基于可逆链终止物和合成测序的 Solexa 及 HiSeq 测序技术，基于离子敏感场效应晶体管检测的 Ion Torrent 测序技术，基于连接酶和简并探针的 PSTAR 测序技术等。虽然这些平台的化学原理各异，包括边合成边测序、边连接边测序等，但它们具有一些类似的样品处理步骤，包括 DNA 片段化，还可以适当改动流程来控制配对标签间的距离，连接平台特异性的反应接头以建立待测片段文库，均有体外扩增过程，包括乳液 PCR（emulsion PCR）或桥式 PCR（bridge PCR）等方法，并分别依赖这些方法使文库单一分子扩增至阵列上固定空间的克隆簇，测序过程是对高密度 DNA 阵列进行酶法操作和荧光或化学发光图像采集（也可以检测半导体芯片检测 DNA 合成过程中释放的氢离子浓度变化）的迭代循环，其生化反应的实现手段各异，但都依赖于聚合酶或连接酶合成 DNA，产生引物延伸系列，最终获得原始测序数据。一般的测序反应是从一端对片断文库进行单端测序（single-end sequencing），即可获得核酸碱基或拷贝数信息，大部分 NGS 平台也可以从两端对片断文库进行两端测序（paired-end sequencing，PE），从而增加测序数据量，提高核酸序列拼接的准确性，并可以发现插入或缺失（indel）以及倒位（inversion）等结构重排变异。此外还可以采用在双向测序基础上进一步优化的配对测序（mate-pair sequencing，MP），以进一步增加对核酸结构变异的识别能力。

### 6.1.4 单分子测序

单分子测序技术被认为是第三代测序技术。其中 SMRT 技术利用荧光信号进行测序，而纳米孔单分子测序技术利用不同碱基产生的电信号进行测序。Pacific Biosciences 公司的 SMRT 技术基于边合成边测序的思想，以 SMRT 芯片为测序载体进行测序反应。SMRT 芯片是一种带有很多零级波导（zero-mode waveguides，ZMW）孔的厚度为 100 nm 的金属片。将 DNA 聚合酶、待测序列和不同荧光标记的 dNTP 放入 ZMW 孔的底部，进行合成反应。与其他技术不同

的是，荧光标记的位置是磷酸基团而不是碱基。当一个 dNTP 被添加到合成链上的同时，它会进入 ZMW 孔的荧光信号检测区并在激光束的激发下发出荧光，根据荧光的种类就可以判定 dNTP 的种类。此外由于 dNTP 在荧光信号检测区停留的时间（毫秒级）与它进入和离开的时间（ 微秒级） 相比会很长，所以信号强度会很大。其它未参与合成的 dNTP 由于没进入荧光型号检测区而不会发出荧光。在下一个 dNTP 被添加到合成链之前，这个 dNTP 的磷酸基团会被氟聚合物（fluoropolymer）切割并释放，荧光分子离开荧光信号检测区。SMRT 技术测序速度很快，利用这种技术测序速度可以达到每秒 10 个 dNTP。而英国牛津纳米孔技术公司的 GridION 和 MinION 测序仪采用了纳米孔测序技术：采用一种特殊的 α 溶血素蛋白七聚体整合进磷脂双分子层而形成出纳米级小洞或小孔。在膜的一侧施加电位差将 DNA 单链（带负电）拉进纳米孔，当 DNA 的不同碱基通过时，引起细微的电流变化，即可识别出不同的碱基序列。纳米孔单分子测序技术还能够直接读取甲基化的胞嘧啶，而不像传统方法那样必须要用亚硫酸氢盐（bisulfite）处理，这对于在基因组水平研究表观遗传相关现象提供很大的帮助。但到目前为止，单分子测序技术的准确性还没有得到明显突破，因此未能大规模推广应用。

### 6.2 测序技术的应用

测序技术被广泛应用于核酸序列的分析。最早常应用于人类疾病的遗传学检测，特别是感兴趣的靶向基因测序。后来被广泛用于各种人类核基因组、线粒体基因组、以及 RNA 转录产物的检测分析，应用领域包括肿瘤诊断和预后分析、药物基因组学检测等。此外，测序技术还被用于检测微生物的基因序列，以检测是否存在耐药基因、决定最佳抗病毒治疗方案等。随着 NGS 的发展，测序技术的应用范畴得到进一步拓展，可以用于检测 Sanger 测序难以检测的多样品检测、多基因组合检测、低含量基因突变检测等。

#### 6.2.1 胚系突变的遗传学检测

遗传检测是指针对人类染色体、DNA、RNA、基因、和/或基因产物进行的检测分析。常用于确定单基因遗传病的致病基因、复杂疾病的易感风险、药物敏感性和毒副作用等。虽然 NGS 已经开始应用于临床，但 Sanger 测序一直是序列

分析技术的金标准。

1）靶向区域的基因测序

靶向区域的基因测序是仅针对基因组中的少数基因、或期望分析的关键位点所在区域的 DNA 片段进行测序，可检测生殖细胞和体细胞突变，分析与人类疾病相关的基因突变。

针对某个单一基因的测序通常采用 Sanger 测序完成，用于鉴定遗传病的致病基因，也可以用于对 NGS 测序结果的准确性和重现性进行确认。

针对多基因组合进行测序（multigene panel sequencing）是指针对与某种疾病或临床症候群相关的一组基因进行测序分析。模板制备时多采用选择性富集的方法，例如，与色素性视网膜炎相关的基因有 100 多个，把这些基因从基因组 DNA 中捕获或扩增出来之后，再进行测序分析。针对多基因组合的 NGS 检测项目在研发和确认时需要大量的经费和时间投入，但可以大大降低后续的多基因检测成本。

2）大范围测序

大范围测序是指对于大的基因片段、外显子组、全基因组进行分析，适用于相关的基因位于多个基因组区域内的疾病。人类全基因组测序（Whole genome sequencing）是指利用新一代测序平台对人的不同个体或群体进行全基因组测序，并在个体或群体水平上进行生物信息分析。通过全基因组测序可获得个体基因组所有的遗传信息，目前能够检测到的遗传变异包括 SNP（single nucleotide polymorphism, 单核苷酸多态性）、Indel（Insertion or deletion, 插入或缺失）、SV(structure variation, 结构变异)等。除了可用于在群体水平上研究物种的进化，环境适应性及自然选择等方面,还有助于快速发现与重要临床表型相关的遗传变异，用于分析人疾病易感性及其他遗传特性，指导个体化治疗等。帮助人们从分子水平进行疾病的诊断、预防和治疗，将大规模基因检测技术转化应用于临床诊疗实践中。

全基因组测序分为两大类：从头测序和重测序。从头测序（*de novo* sequencing）是指不需要任何参考序列对某个物种进行测序，再通过生物信息学分析方法进行拼接、组装，从而获得基因组的序列图谱。而重测序（re-sequencing）

是指在物种基因组序列已知的条件下，对不同个体进行基因组测序。因为通过人类基因组计划已经获得人类的基因组序列，因此，目前针对人类的基因组测序都属于重测序。

而针对编码蛋白质的外显子组进行测序在临床中更为常用，人类外显子包含大约 2-2.5 万个基因，大约有 18 万个外显子，占人类基因组的 1%-1.5%。

3）染色体非整倍性检测

染色体非整倍体疾病主要指染色体在数目或结构上的改变，染色体整组或整条的增减，可使细胞的遗传功能受到损害，扰乱基因之间的平衡，影响物质代谢的正常过程，造成多器官、多系统的畸变和功能改变。临床中比较常见的是染色体三体综合征，即细胞内某染色体的数目不是正常的两条而是三条，包括最常见的：唐氏综合征（T21）、爱德华氏综合征（T18）和帕陶氏综合征（T13）等。

自从 Lo YM 等于 1997 年证明了孕妇外周血中存在胎儿游离 DNA（cell-free fetal DNA，cffDNA）以来，开启了基于 cffDNA 的非侵入性产前筛查（Non-invasive prenatal diagnosis，NIPD）方法的研究。胎儿 cffDNA 来源于胎盘凋亡的滋养层细胞，经过胎盘屏障进入母血，怀孕 4 周左右就可以从孕妇外周血中检出 cffDNA，孕 8 周建立胎盘循环后，cffDNA 以相对固定的比例（5%-10%）稳定存在于母体外周血血浆中。随着测序技术的发展和应用，目前已经可以采用新一代测序技术对孕妇血液游离 DNA 进行深度测序、最终通过分析胎儿 DNA 片断占正常母体 DNA 的比例，来分析胎儿染色体数目异常导致的母体血浆中 cffDNA 含量的微量变化，判断胎儿是否存在染色体非整倍性。

### 6.2.2 群体测序

群体测序是指针对捕获片段或靶向扩增后的扩增子（amplicon）进行测序，以发现基因组背景复杂的、异质性（heterogeneity）样品中低含量的突变。Sanger 测序可以检出突变比例为 15%-25%的突变，而 NGS 的检测通过将相同区域的大量读段进行比较，检测灵敏度可以达到 1%-5%。

这种用于检测异质性样品中的低含量突变的测序技术，目前主要被用于肿瘤、感染性疾病、宏基因组学和线粒体突变分析。

（1）肿瘤

对于来自肿瘤患者样品的核酸定量、定性和纯度都存在问题。因为除了少部分肿瘤穿刺样品外，大部分肿瘤组织是一个由正常细胞和肿瘤细胞混合的异质性细胞群体。因此，检测出来的体细胞突变比例大都小于 50%。此外，经过处理后的 FFPE 标本中的核酸的降解比较明显，大部分肿瘤来源的核酸片段都是碎片化的，长度小于 150bp，而且和组蛋白之间的交联很难打开。

NGS 目前已广泛用于检测肿瘤样品中的错配、插入/缺失、拷贝数变异、染色体重排等突变类型，有助于阐明癌症机理，帮助临床医生进行癌症的诊断、治疗指导和预后分析等。

（2）感染性疾病

在感染性疾病中，不需要培养和克隆扩增，就可以采用群体测序对某一种或某一群病原体进行检测。用于鉴定特定的耐药细菌。帮助临床选择合适的抗生素。对于一些体外很难培养或生长很慢的病原体（例如结核杆菌），基因测序的价值在诊断和用药指导方面的价值都很明显。

在感染性疾病暴发流行时，还可以通过全基因组测序对病原体进行鉴定。

（3）宏基因组学

宏基因组检测是指对于特定来源的样品中的全部病原体的基因序列进行检测和分析。通常可以检测 16sRNA 的 V2 和 V6 可变区进行成份分析。

（4）线粒体基因检测

线粒体 DNA 突变可以导致母系遗传性疾病。每个人细胞内的线粒体基本上都是存在一定程度异质性的混合群体（由遗传来源不同的线粒体群体组成），只是采用常规的测序技术不易检测到较低比例的线粒体基因突变，而采用高覆盖深度的 NGS 技术、则可以更灵敏地检测到低含量的线粒体基因突变。

### 6.2.3 RNA 测序

人体的转录组是一个混合体，包含各种编码和非编码 RNA，如 mRNA，microRNA，RNA 前体等。采用 NGS 对 cDNA 和 RNA 等进行测序，除了分析编

码基因和 microRNA 的表达变化、在 RNA 水平发现突变以外，还可以发现各种可变剪接、融合基因变异体等，在疾病的病因诊断、治疗药物选择等方面均发挥重要作用。

### 6.2.4 甲基化检测

DNA 中胞嘧啶（C）的甲基化是真核生物表观遗传学调控的重要机制。例如，对于基因组中 DNA 甲基化谱进行分析有利于癌症发生发展机制的理解。采用 NGS 方法，可以将甲基化分析范围由原先的几个位点扩展到全基因组水平、而且精度达到单碱基水平，即为甲基化组分析。

### 6.2.5 染色质构象分析

除了通过三联密码决定蛋白质功能之外，不同区域染色体的相互作用也可以在基因表达、基因组稳定性方面发挥作用。采用“染色体构象捕获（chromosome conformation capture）”和 NGS 相结合的方法，可以对染色体构象进行分析。

### 6.2.6 大片段的结构变异

人类遗传学变异包括很多类型，小到单一碱基改变，大到大片段染色体水平的变异——即结构变异（structural variations，SV）。已知结构变异与许多遗传性疾病和复杂疾病的易感性、发生、发展相关。NGS 方法为染色体结构变异研究提供了革命性的解决方案。与传统的芯片分析方法相比，NGS 最重要的优点就是通过一次实验就可以发现多种结构变异。但是，采用 NGS 进行结构变异的生物信息学分析工作仍有很多挑战，特别是大片段的结构变异，目前的拼接算法还有很多局限性。

## 7. 样品处理

恰当的样品处理是确保样品完整性和核酸定性定量检测准确性的关键环节。样品在检测前必须确保样品的采集、运输和储存符合要求。样品处置不当可能引起核酸降解，导致基因测序失败或检测结果不准确。主要环节包括：

a)确保患者样品的采集方法正确。

b)在采集过程中确保样品和信息的完整和准确性。

c)在检测前和检测后的样品的运输和储存过程符合规范。

d)患者样品处理（例如核酸提取）符合规范。

这些环节适用于所有采用测序方法进行人或微生物基因序列测序的临床检测项目。

## 7.1 样品采集、运送和保存

### 7.1.1 信息采集

样品采集前需填写送检申请表，提供受检者必须的信息，为医务人员进行适当的项目检测和采取治疗措施提供参考。需采集的信息包括：

a) 常规信息：包括样品唯一性编号或条码、采样日期、采样时间、受检者姓名、性别、出生日期、样品来源（所采样品的组织类型）、采样单位、采样人姓名。

b) 根据检测项目制定有利于临床医学指导的临床信息采集表，信息表包含的内容包括检测项目、样品唯一性编号或条码、受检者姓名、出生日期、年龄、民族、采样日期、样品类型、相关的临床资料（如身高、体重、疾病诊断、疾病分型分期、合并疾病、用药情况）和临检信息、采样单位及科室名称、医生姓名、送检目的等信息。

c) 临检实验室应有专业人员根据信息采集表对检测项目的合理性进行审核，必要时可与送检医生讨论。

### 7.1.2 患者的正确识别及知情同意

患者的正确识别是确保获得正确的临床样品的前提。收集样品的容器上应注明患者的唯一信息，通常应包括检测条码或编号、待检者的姓名、送检科室和住院号等信息。医护人员在采样前需首先核对确定患者的身份，核实患者的姓名、性别、住院号等能标示患者的信息。

样品收集前应向患者讲解基因检测的意义，以得到患者的认同，即知情同意。对于涉及遗传基因信息的临床检测项目，所有受检者均需签署知情同意书，告知所检测项目的目的、意义、基本过程、剩余核酸的去向及保存时间、临检样

品是否可匿名用于科研项目等，确保受检者的个人隐私（包括医疗记录和医疗数据）得到保护。

对实施有创检查的分子诊断项目如穿刺取活检组织，采集人员应首先对检查可能遇到的风险清楚地告知患者及家属，并对紧急情况下的紧急预案如实告知，使患者及家属全面了解某些特殊检查可能带来的后果。知情同意书是医方履行如实告知义务的证据，也是患者行使选择权的书面依据。

### 7.1.3 样品的采集、运送和保存

可以用于测序的样品有很多种，包括全血样本、血浆样本、组织标本（新鲜组织、冰冻组织、石蜡包埋组织、穿刺标本）、口腔拭子和骨髓等。为确保样品采集的质量，避免污染和干扰，负责采集样品的临床医生需进行样品采集要求培训。无论采集哪种类型的样品，采样时都必须戴手套，这样既可避免样品中病原微生物感染，又可防止采样人员的皮肤脱落细胞污染样品。临检实验室应向样品采集和运输人员提出样品收集、处理、运送和保存过程合适的条件要求。

各种样品的采样过程要遵守卫生部《微生物和生物医学实验室生物安全通用准则》和《个体化医学检测质量保证指南》中关于“样本的采集、运送和保存”的要求。

## 7.2 核酸提取方法及质控

1） DNA 提取用酚-氯仿提取法和盐析法等均可。酚氯仿法提取的 DNA 可能导致 DNA 样品中酚或氯仿残留，从而抑制后续的 PCR 反应。盐析法提取的 DNA 可能存在蛋白质及其他物质的残余，DNA 的纯度和得率不高。DNA 提取要求在生物安全柜内进行操作。DNA 要求 OD 值介于 1.6-1.8 之间，浓度大于 50 ng/μL。

2） DNA 相对稳定，在无 DNA 酶的情况下，常温下纯化的 DNA 在 TE (Tris-EDTA) buffer 可放置 26 周，2-8 ℃ 冰箱中可放置至少 1 年。为降低 DNA 酶的活性，确保 DNA 的完整性，长期保存纯化的 DNA 样品应在 0 $^{o}$C 以下的环境中。DNA 应放置在带盖密封、疏水的塑料管中（带橡胶垫片的塑料管更好，可防蒸发）。聚丙烯容易吸附 DNA，尤其是在高离子强度的时候；

聚乙烯结合 DNA 的能力更强。DNA 最适于保存在异质同晶聚合物材料的塑料管或经特殊处理的聚丙烯塑料管中。DNA 一般溶解在 pH 为 7.2 的 TE 溶液中，可减少 DNA 的降解。但如果 DNA 在提取后几天内用于 PCR 或酶切目的，也可用双蒸水进行溶解。建议将 DNA 原液保存于-70$^{o}$C 或以下的环境中。当同一个受检者的 DNA 样品要进行多个检测时，建议将 DNA 样品分装后保存，这样既可以减少反复冻融引起的 DNA 降解，又可减少样品间的污染。

3） 血浆游离 DNA 的提取过程中需要尽可能消除血浆中各种可能抑制 DNA 聚合反应的抑制成分，包括血浆蛋白、血红蛋白、细胞碎片等，建议采用基于微柱吸附技术的商业化提取试剂盒。根据所需检测的项目不同，提取的游离 DNA 总量应为 50 ng-1000 ng 因为血浆游离 DNA 浓度很低，需要采用实时荧光定量 PCR 技术进行定量。对于纯化后的游离 DNA，应根据实验项目对于 DNA 量的需求、及时分装，暂时不用于检测的 DNA 标本应及时冻存于-80 $^{o}$C 或更低的温度条件下，避免反复冻融而加剧游离 DNA 的碎片化程度、降低后续检测的灵敏度。

### 7.3 检测后样品的保存和处理

样品在检测后要进行一定时间的（尽可能长期）保留，以备必要时复查。样品的保存也可为科研工作的开展和回顾性调查提供条件。完成检测后剩余的 DNA 样品至少在-80℃保存 2 年。DNA 在-70℃的环境下可保存至少 7 年。纯度不高的 DNA 样品建议保存在-20℃或更低的温度中，以确保 DNA 的完整性。在不影响受检者个人隐私及利益的前提条件下，DNA 及临床资料可以匿名用于科学研究。废弃的样品应作为生物危险品处置。

## 8. 测序模板制备

对于特定区域的核酸片断进行成功测序的关键，是需要待检测区域的核酸片断具有足够的质量（quality）和数量（quantity）。Sanger 测序和 NGS 的关键步骤参见下图：

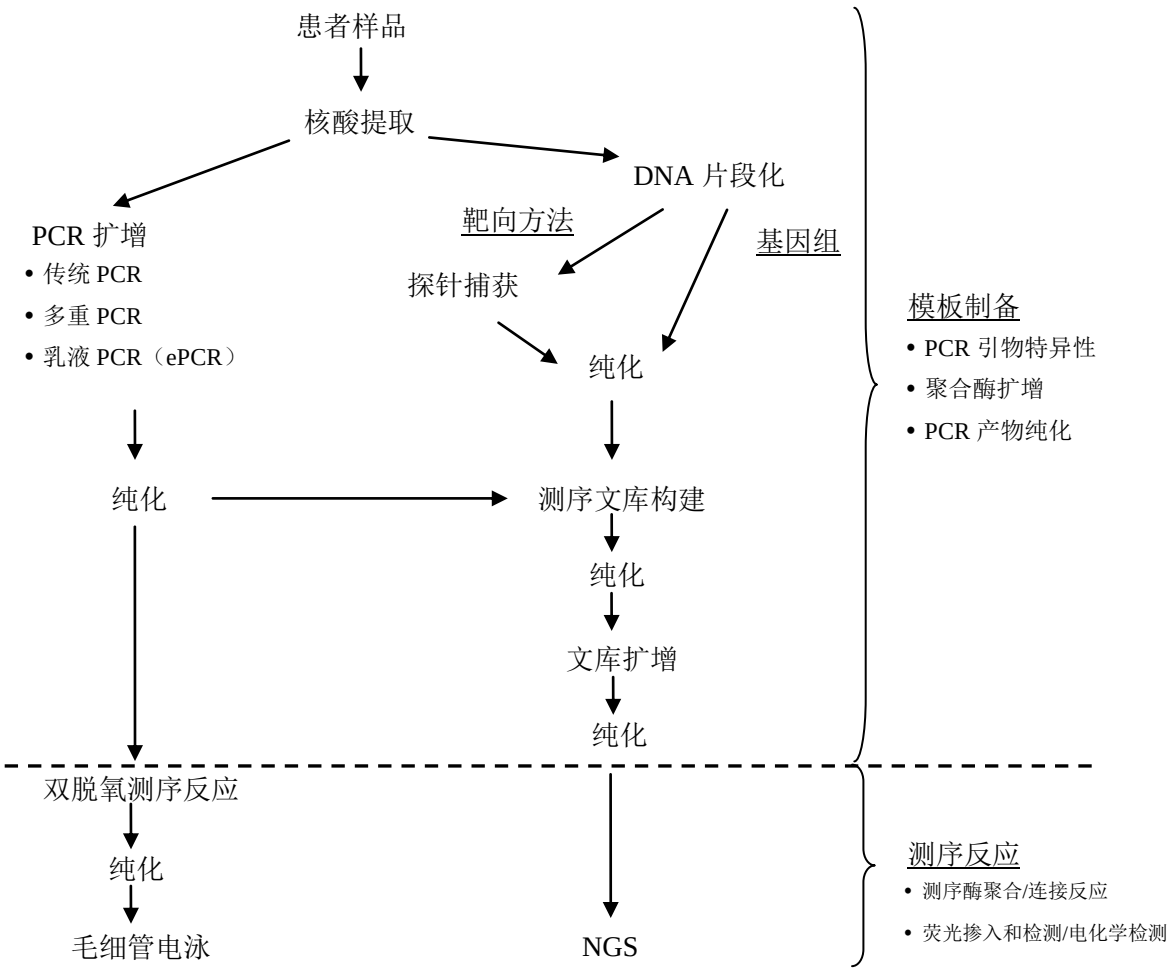


图 1 DNA 测序前的模板制备步骤

DNA 提取后，可以通过直接或靶向捕获的方法进行测序文库构建。PCR 和探针捕获技术均可用于靶向测序区域的分离，每种方法均可用于下游的文库扩增和测序过程。如果靶向测序区域的数目很多时，主要考虑探针捕获技术，例如外显子组测序就是采用探针对人类基因组中所有编码基因的外显子及其侧翼序列进行捕获。但如果靶向序列的拷贝数较低时，则需要采用 PCR 扩增技术进行扩增和分离。在 Sanger 测序中，PCR 获得的待测片段还需要进一步纯化才能用于测序反应，双脱氧测序反应和毛细管电泳检测是分别进行的；而在 NGS 中，还包括下游的文库构建、扩增和纯化等步骤才能上机测序，而且 NGS 的测序反应和原始测序数据的信号检测是同时进行的，因此又被称为“实时（real-time）”测序。

### 8.1 Sanger 测序的模板制备

当然，随着测序技术的不断发展，关于模板准备过程中的一些关键需求可能也会随之改变，但一般来说目前还不可能直接对来自患者的样品进行测序。PCR 技术是一项可以用来分离和扩增待测区域核酸片段的重要手段之一。因 PCR 不需要进行繁复的克隆和细菌文库扩增操作，就可以对特定的核酸序列进行扩增，已经被广泛用于 Sanger 测序和 NGS 的模板制备中。对于 PCR 产物进行成功测序有以下基本要素：

a）引物的设计和选择

b）扩增反应体系的配制

c）PCR 循环的扩增参数

d）控制 PCR 产物的气溶胶污染

每一个要素都需要进行优化才能保证获得足够的测序模板、同时减少非特异性扩增。

用于 Sanger 测序的 DNA 模板要求进行光密度检测以分析其质量和浓度。一般要求 A260/A280 比值大于 1.8 以上，A260/A230 比值大于 2.0 以上。

用于 Sanger 测序反应的模板浓度太高时，会导致初始的反应峰太高而之后的信号迅速衰减，如果模板太少，则会导致峰高和信噪比明显降低。测序模板 DNA 的总量要求与特定检测项目的 PCR 产物大小、模板是双链还是单链等有关。一般来说，对于 100-200bp 的 PCR 产物进行测序，需要 1-3ng 模板；而对于 1000-2000bp 的 PCR 产物，至少需要 10-40ng 模板。

### 8.2 大规模平行测序（NGS）的测序文库构建

#### 8.2.1 文库构建概述

新一代测序的模板制备过程与经典的 Sanger 测序有所不同，首先、也是最重要的一步就是构建文库（library），主要是由连接了与测序平台相匹配的各种寡核苷酸接头（adapter）的 DNA 片段组成。应用于 NGS 的文库构建主要包括 3 个步骤（参见图 1）：1）片段化；2）富集；3）克隆生成；通常还需要经过一步

纯化步骤。用于构建文库的核酸序列可以是基因组 DNA（gDNA）、常规 PCR 产物（PCR amplicon）或由 RNA 反转录而来的 cDNA。模板 DNA 的质量要求同 Sanger 测序。用于构建文库的核酸经过超声、雾化、酶切等步骤完成片段化过程，随后进行末端修复、磷酸化修饰、再连接上与测序平台相匹配的接头，即可完成文库构建。一般连接后的产物需要电泳分离、并选择片段长度与测序平台相匹配的产物用于下游分析。在部分平台，还需要进一步采用与接头互补的引物进行 PCR 扩增以提高文库的浓度。

对于已经构建的文库，还需要综合采用基于荧光定量 PCR、荧光计、毛细管电泳等技术的检测平台，对文库质量进行定量和定性评估。

### 8.2.2 DNA 片段化和连接

构建文库之前需要对 gDNA 进行片段化。可以用于 DNA 片段化的方法包括超声、雾化、酶切等。建库前的 gDNA 需要事先采用 A260/A280 或荧光染料进行浓度和纯度分析，进行凝胶电泳分析是否已经发生降解，DNA 电泳时条带已经明显弥散（smearing）、或者呈现凋亡特征性的梯形 DNA 片段特征，都不适于再进行下游处理。

### 8.2.3 富集

对于很多临床检测项目，并不需要检测全基因组序列，而只需要对于临床相关的靶向区域进行富集和针对性测序即可。对富集后的区域进行重测序除了结余时间和成本之外，也便于下游的测序数据分析和解读。

#### 8.2.3.1 靶向区域扩增

采用 PCR 的方法，对感兴趣区域的基因组区域进行扩增富集。采用本方法最大的缺陷是可能导致二倍体样品的两个等位基因的不平衡扩增，这种偏倚的产生原因主要与扩增引物的结合位点有关。其他问题还包括对于高 GC 含量区域或导致扩增失败。因此需要在建立检测项目时对引物设计进行优化和校验。

#### 8.2.3.2 靶向区域杂交捕获

通过固相或者液相的杂交探针，把感兴趣区域的基因组片段捕获下来，再

进行下游的检测与分析。本方法的最大优点是易于使用。而且合成和优化探针都有比较成熟的商业化公司提供服务。对于本方法的最大的需求就是探针能够覆盖全部感兴趣的基因组区域，但一些富含的 GC 区域可能很难捕获，而重复区域又会捕获的太多。由捕获导致的测序不均衡需要在测序中增加覆盖深度，才能保证那些捕获的较少的区域能够达到最低检测灵敏度而被检测出来。

#### 8.2.3.3 非靶向性富集

如果 NGS 检测的目的不是局限在某个特定区域、而是想分析特定长度的 DNA 片段，可以通过剪切破碎后凝胶电泳的方式对 DNA 或 cDNA 样品进行分离，然后将所需要分析大小的片段回收、连接接头，进行后续反应。采用本策略进行特定长度片段富集时需注意：1）必须对 DNA 剪切破碎方法进行校准，以确保预期的大小片段的最大分布，可以根据 DNA 质量（或是来源于 gDNA 还是 mRNA）进行调整。2）如果采用从凝胶中切除 DNA 的方法,必须用无菌的或一次性的切除工具，而且建议把切口的位置和宽度用照片的形式记录在案。而且纯化也至关重要，因为任何残留的凝胶或其它提取试剂可以影响后续的各步反应。3）对于回收后的合适大小的 DNA 进行处理和后续的寡核苷酸连接时也需要非常小心，注意勤换手套、尽可能避免污染。

#### 8.2.3.4 全基因组和全转录组扩增

如果对于起始量很小的生物学材料进行基因组测序时，例如循环肿瘤细胞、母体血液中的胎儿细胞、循环中血液中的游离 DNA、细针穿刺组织标本、激光显微切割获得的少量组织标本、其他单细胞分析等，就需要采用全基因组或全转录组扩增的方法对标送检本进行扩增富集，以确保痕量的起始核酸也能够被扩增达到 NGS 的最低检测灵敏度。对全基因组扩增过程中可能产生的扩增错误或模板偏倚，需要采用 SNP 或 CGH 等方法进行校正。

### 8.2.4 克隆生成

#### 8.2.4.1 乳液 PCR（ePCR）反应

乳液 PCR 需要先将连接了接头的片段化文库杂交到微球（bead）上，DNA 片段通过与微球表面上和接头互补的序列结合而紧密相连。乳液 PCR 是将每一

个连接了一种DNA片段的微球分配到一个油包水的乳滴中，然后进行PCR扩增，PCR 完成后，破碎乳液即可得到表面包被克隆和富集的文库的微球，将其均匀分散在微孔板（边合成边测序所用的 PicoTiter™Plate）或玻片反应池中（用于连接测序（SBL）的寡核苷酸探针）。分散时要求每一个微球在一个孔中或与一个寡核苷酸探针结合。通过这种克隆扩增方法，可以保证微球表面具有足够多的DNA，从而在下一步的测序反应中产生足够检测的光学或电化学信号。

#### 8.2.4.2 固相成簇扩增

对于采用循环可逆性末端终止（cyclic reversible termination）的测序平台，将连接了接头的片段化 DNA 稀释后自动散到玻片反应池、与表面的单链引物碱基互补，从而被固定于芯片上，另一端与附近另一个引物随机进行互补结合，从而形成“桥”，进行 30 个左右循环的等温桥接扩增（bridge amplification）技术进行扩增，最终，每个结合到单一表面的 DNA 分子会被放大 1000 倍以上，从而成为单克隆 DNA 簇，可以用于下一步的测序反应、产生读段（read）。

### 8.3 特殊测序模板准备时的注意事项

#### 8.3.1 宏基因组学分析

宏基因组学分析需要直接从特定微环境样品中提取全部微生物的 DNA，而不同微生物的核酸提取方法可能各异。因此，在选择核酸提取方法时需将由此产生的方法学偏倚控制到最低，尽可能选择可以将特定患者样品中全部微生物核酸的方法。

对于采用测序技术进行病原微生物（例如细菌、病毒和霉菌）鉴定，经常遇到的难题是如何检出样品中的痕量的微生物，因此，必要时需要事先进行靶向捕获和富集技术以提高检测灵敏度。在开发项目过程中，需要针对样品制备的检出限（LOD）进行专门优化和标准化。

此外，微生物核酸在环境中也无处不在，因此，在样品制备和处理的全过程中，均需要设立可靠的对照实验，以避免外源性微生物核酸混入待测样品中干扰检测结果。

#### 8.3.2 RNA 相关文库

对于用于新一代测序的 RNA 模板检测，除了严格遵循 RNA 实验的注意事项之外，还需要在文库构建之前再对 RNA 质量进行检测和评价。而且在提取时还需要加上 DNA 酶处理步骤以去除所有 DNA。建议根据实验需求，采用专门针对总 RNA、mRNA、microRNA 的提取和处理的商业化试剂盒。

### 8.4 采用 NGS 进行多样品混合检测（sample multiplexing）

因为 NGS 巨大的单次检测通量，可以将多个样品混合在一个测序泳道或反应池中进行同步检测。NGS 通常采用在每一种样品的测序文库中加上唯一的寡核苷酸标签的方法进行多样品检测，这些寡核苷酸标签由“条码（barcode）”，在不同平台中也被称为“标签（index）”组成，在测序时与待测 DNA 片段同时被检测出来，从而产生一个可用于区分不同样品的唯一编码。通过这种方法，可以极大降低每单一样品的测序成本。

多样品混合检测的实施通常是在文库构建过程中，在所有 DNA 分子测序接头的 5’或 3’端，加入一段长 6-8bp、序列已知的短链 DNA 作为“条码”或“标签”。测序结束后，通过软件分析和筛选，即可从混合测序的结果中获得每个不同样品的测序结果。

采用这种方法测序时，在设计实验后分析结果需要注意可能存在的问题：1）条码检出率的不均一（特别是采用嵌入式/in-line 条码技术时）；2）标签互换（index swapping）：因为测序数据读取错误所致的条码关联错误，一个患者的测序数据可能被关联到另外一个患者，可能的发生原因包括在多重检测样品制备过程中的“条码”或“标签”相互污染、存在干扰样品、聚合酶保真性不足等，当然在测序和测序后分析过程也可能发生。因此，在临床检测项目中，需要采取各种实验条件优化措施，把标签互换比率（index swap rates）降到最低；一般要求标签交换率要比整个分析的灵敏度低几个数量级（例如 0.01%），特别是针对低含量突变的检测，更需要尽可能降低标签交换率以确保不能影响监测结果的准确性。

## 9. 测序步骤与可能存在的问题

### 9.1 检测方法概述

本节将重点关注测序过程中需要注意的特殊步骤以及由此可能产生的问题。

这些问题可能是在各种测序平台和技术中都普遍存在。

Sanger 测序技术已经相当成熟，可能产生的问题也比较清楚，相关的质量管理和控制重点和措施已经被许多临床实验室采用。由于该方法可直接读取 DNA 的序列，因此是被认为是基因分型的金标准。Sanger 测序法的操作过程主要包括 PCR 扩增和 PCR 产物纯化、测序反应、测序和结果分析四个主要步骤。分析时需要设置阴性对照和阳性质控品，当阳性质控品没有出峰时提示实验失败，确认 DNA 质量好后，采用同批号试剂和同一台仪器重复实验，并确保检测试剂是否按要求保存。当阴性对照品出峰时，说明有污染，需要找出污染源后重新进行实验。该方法属于定性检测，优点是测序长度较长，可发现新的变异位点。主要不足：灵敏度不高，尤其是在进行肿瘤组织体细胞突变检测时，当组织中靶标基因突变比例低于 20%时，可能出现假阴性的结果；对试剂和仪器有特殊要求，不易普及；操作复杂，成本相对较高，速度慢、通量低。

但 NGS 测序技术发展很快，相关的标准和质量参数很少、甚至没有，因此，对于临床实验室使用时就需要特别注意遵照本指南的相关条款，对于测序设备及数据分析软件的性能、分析结果的确认等进行评价，才能最终获得可重复、高质量的测序结果。而且，还需要掌握特定测序设备的性能和规格，熟悉其在用于临床检测和分析中的优点、缺点和特殊要求，特别是在样品类型、模板质量、序列混合方式等方面。

**9.2 测序方法和仪器的选择**

NGS 技术正在迅速发展。最初的 NGS 检测平台主要是为了科研所需的大规模检测通量而设置，随着技术进步以及面向临床的简单、快捷测序需求的增加，又有多种小型化的台式测序仪推出。据预测，这些小型化的测序仪还将不断推陈出新，而且价格会更便宜、性能也更优越。但正是因为有多种可供选用的技术平台，而且每一个都有自己的优点和缺点，必须以动态的眼光分析现有的技术平台、根据实验室的需求进行权衡选择。

下列的七个主要因素可供在选择测序平台时权衡考虑。

### 9.2.1 测序通量

测序通量表示仪器在单次运行时可以产生多少的数据量。高测序通量可以满足全基因组测序的需求，或者一次获得更大的覆盖深度或检测更多的混合样品。主要的缺点是单次运行成本较高，而且所需的计算资源明显高于中等或低通量的测序仪。如果样品数量不多、或仅需要对靶向区域进行测序，较低通量的系统（甚至 Sanger 测序）就可以满足需求。

### 9.2.2 样品通量

样品通量是指在给定的时间内可以检测的样品数目，将由仪器运行时间，测序通量，多样品混合检测容量，以及预期的应用对象等确定。

### 9.2.3 读段长度（读长）

读段长度是指在单次测序反应产生的碱基的数目。更长的读长可以简化序列比对过程，并可以较为准确地判断特定区域的单倍型。读长的增加将直接延长仪器运行时间、以及单次运行的成本。较短的读长可能在变异检测方面存在局限性。这些属性权重的高低，将取决于检测项目的范围和性质。

### 9.2.4 覆盖深度

覆盖深度是可以用于特定区域碱基识别的独立的读段数目。针对某一区域的覆盖深度增加时，最终拼接完成的共有序列出错率就会下降。

由于目前 NGS 平台采用的技术原理各不相同，最终获得的测序读段长度也长短不一；因此，在测序数据准确性相同的条件下，进行重测序时所需覆盖深度也各不相同。不同测序读长和对应的参考覆盖深度要求如下表 1 所示：

表 1 不同读长的平台进行基因重测序时所需要的覆盖深度*

| 平均测序读段长度（bp） | 重测序所需的覆盖深度 | 参考测序平台 |
|---|---|---|
| -50 | 1000 | CG，Illumina/GA，HYK/PSTAR-IIA |
| 51-100 | 50-100 | Illumina/HiSeq |
| 101-300 | 30-50 | Ion Torrent |
| 301-400 | 10-30 | 454 |
| 401-500 | 8-15 | |
| 501-600 | 6-10 | Sanger |
| 601-1000 | -5 | |

*推荐的覆盖深度参考 NHGRI 的数据：http://www.genome.gov/sequencingcosts/

此外，特定基因组区域所需的覆盖深度可受到序列结构的影响，例如有些富含 GC 区域、或碱基重复区域，可能需要更深的覆盖倍数才能产出质量合格的序列。

针对不同的检测项目需要的具体要求，也将影响测序通量和多样品混合检测容量。为特定临床项目建立覆盖深度标准时，应考虑到该项目所需的分析准确度和精密度。如果对遗传背景不同的混合样品进行测序时，就需要更高的覆盖率才能成功检出变异。例如，同样是读长 150 bp 的测序平台，如果检测外周血有核细胞基因组 DNA 的胚系突变，需要 50-80 倍覆盖深度，如果用于检测来自 FFPE 标本的肿瘤细胞体细胞突变，则需要 500-1000 倍的覆盖深度。

### 9.2.5 成本

NGS 的检测成本包括初始测序设备的成本，单次运行的试剂成本，以及下游的数据分析和生物信息学费用，应根据实验室的预计检测通量和后续的潜在增长进行评估。尽管供应商可能会降低试剂的费用，但在大多数情况下，检测通量高的仪器的单次运行成本也高，因为试剂的消耗是按照每次运行计算的，而与每次运行时检测了多少个样品无关。因此，必须要让仪器运行接近满负荷，才能更好地控制单个样品的平均检测成本。另外一个需要考虑的是测序仪维护的复杂度和所需要的时间成本。此外，NGS 测序仪需要配备训练有素的实验技术人员，从事个体化医学检测的人员都必须经国家规定的相关培训并取得合格证书，这都是需要考虑的人力成本投入。

### 9.2.6 运行时间

运行时间指单次运行生成数据所需的时间。通常依赖于读段长度和生成的数据量。仪器的测序通量越高、运行时间越长。但是，增加的测序通量也可以用于混合样品的检测。如果临床应用检测项目可以接受较长的运行时间，通过增加测序通量就可以降低单个样品的平均成本。检测通量低的测序平台一般运行时间较短，可以快速返回检测结果。其他需要考虑的时间因素还包括样品制备、文库构建以及数据处理，这些都可能显著增加测序项目完成所需的总体时间。

### 9.2.7 测序首次成功率

此外，在检测项目设立和优化时还需注意，测序的首次成功率也应该是临检项目中需考虑的重要因素，因为对于大多数 NGS 测序仪，单次运行时间都将花费数小时甚至数天，如果某个项目总是失败、或者需要多次重复检测才能得到可靠的结果，将会大大影响报告发出的时间、并增加实验成本。

### 9.3 测序技术的潜在缺陷或特征可能导致的问题

测序的操作人员必需娴熟掌握和理解检测过程、与临床相关的检测结果的范围，以及测序的技术平台和生物信息学分析软件可能存在的问题。在评价原始测序数据、数据分析过程、质量值（quality score）分析，软件得出的结果解释等过程中也可能会出现问题。在实际检测项目开发和验证中均需要多加注意。

以在各种测序反应中广泛应用的荧光素（fluorophore）为例， 如果不是新鲜配制的荧光素，可能会因为暴露在光线中而快速降解，导致后续测序反应中的检测信号明显衰减；此外，某些荧光素还可能会因为散射而导致临近碱基检测信号“噪音（noise）”增加，甚至识别错误。对于这些因为荧光素造成的测序技术问题，需要通过设立样品间标准化方法、或调节测序仪的检测信号时的灵敏度等进行调节或纠正。

另外，在数据分析时，虽然多个高通量测序平台产出的序列文件均为 FASTQ 格式（其中包含了每个碱基的质量值信息），但是对于每个碱基的质量编码标示，即 ASCII 编码方式，不同平台的不同的软件采用不同的方案，例如：Sanger，Phred 质量值的范围从 0 到 92，对应的 ASCII 码从 33 到 126；Solexa/Illumina 1.0, Solexa/Illumina quality score，值的范围从-5 到 63，对应的 ASCII 码从 59 到 126；Illumina 1.3+，Phred quality score，值的范围从 0 到 62 对应的 ASCII 码从 64 到 126，在实际数据分析时需要详细查考对应平台的帮助手册，这一点在评价原始数据以及进行质量控制时需要特别注意。

### 9.4 碱基识别和质量值

#### 9.4.1 碱基的质量值

454 和其他 NGS 在很大程度上都依赖于碱基质量值来判断测序反应的性能和产出数据的质量。每个测序平台计算均基于仪器的规格以及测序数据生成过程

来计算碱基质量值。尽管计算方法会存在不同，它们都是以对数值的形式来指明碱基的错误率（见表 2），这有时也被称为“Phred-like”质量值。

表 2 碱基质量值与错误率的关系

| 碱基质量值 | 碱基识别错误率 | 碱基识别准确率 |
|---|---|---|
| 10 | 1/10 | 90% |
| 20 | 1/100 | 99% |
| 30 | 1/1000 | 99.9% |
| 40 | 1/10 000 | 99.99% |
| 50 | 1/100 000 | 99.999% |

Phred 值最初采用大的查找表，利用峰图和其他特征来估算 Sanger 测序读段中碱基识别的质量，Phred 值大于 30 规定为高质量。并非所有的软件包使用原来的查找表，但 Sanger 读段的质量值仍称为 Phred 值。Sanger 测序错误率的定量估计限定为 Phred 值；但是，这一指标没有兼顾多个 Sanger 读段的信息联合，也不能作为比对和变异识别的指标。比对的质量通常是由操作者控制的，缺少度量来定义比对的好坏程度。

Phred 最初是一种研究工具，要整合到临床环境很繁琐。现在则有了许多适合应用于临床实验室的商业软件包。这些程序可以来自于测序仪器制造商以及独立供应商，后者开发多种测序平台下有用的数据分析程序。进行诊断序列分析的实验室需要使用这些软件来控制碱基识别的质量。评估每个测序反应所产生序列的平均 Phred 值，可以帮助实验室分辨由于低质量模板，不充分的 PCR 扩增，或整体仪器误差（例如，毛细管阵列问题）形成的低质量序列。

Phred 值和计算公式广泛应用于多个测序平台。这些 Phred-like 质量值，通常称为“Q 值”，也是计算错误的可能性，但通常明确的加入了信噪水平，簇/磁珠重叠（探测器无法解读），碱基掺入率（滞后/流速）等参数。尽管都采用类似 Phred 格式，不同平台的 Q 值不一定是完全等价的，因为每个 NGS 平台之间的底层数据产生和碱基质量算法存在计算差异。当研究这些计算时需要参照仪器和供应商手册；NGS 仪器的常规操作一般不需要修改，但需要跨平台比较时就应当非常谨慎。除了原始碱基识别的质量值，NGS 测序仪还有其他的覆盖深度（coverage depth）、链偏倚（strand bias）等更多指标可用。质量控制应当综合使

用这些质量值和整体指标。

NGS 的读段 Q 值由于测序技术和序列内容的不同可以有很大的变化。一个例子是，在大多数 NGS 平台当出现同聚体，微卫星，插入缺失或者其他类型的序列变异时，Q 值大于 30 的碱基数目将下降，在不同平台下降的程度不同。

另一个例子是，对许多技术而言，“平均 Q 值”是在序列读段的开头是高的，但是随着读段延伸逐渐降低。由于序列读段是独立的，整体的碱基识别质量来自于所有的测序读段。因此这可能导致过高或过低的“整体 Q 值”。Q 值也可以通过重新校正来修饰。这个过程考虑比对质量和增强或减弱特定碱基识别的其他因素。

来自于随机序列和来自目标捕获或基因组特定区域 PCR 扩增的序列 Q 值也需要区分对待。Q 值基于基因组水平的平均值，对于特定目标区域许多程序推荐通过重新校正来得到更准确的数值。Q 值受到每个读段中序列变异（例如同聚体或者微卫星）的影响。对 PCR 扩增产物来说，平均 Q 值依赖于这些因素在扩增子中的长度和位置分布。这并不是说，这些碱基和变异不能被准确识别，但建议仔细检查基因组的目标区域，并设置不同于一般 Q 值的有效性度量水平。

除了碱基的 Q 值，序列读段通常还给出比对质量值。这个值表面该区域比对回参考序列的配对程度。小的变异一般不会影响这个数值；然而，插入缺失，同聚体，和拷贝数等变异会有显著影响。此外还有来自比对方式的影响。启发式比对算法，主要用于大量序列读段的比对，不能很好的处理插入缺失和其他大片段的变异。在这种情况下，通常用改进了的 Smith–Waterman 算法来重新比对。这会增加相当多的分析时间，但需要这么做的区域是可以识别和定位的。插入缺失的局部重新比对，包括同聚体，微卫星，和其他变异，可以提供很大的改进。基因组中假基因或者其他高度同源序列的存在也可以导致错配。在这些情况下，全局比对相较于局部比对可能减少，但不消除问题。

对于存在变异的高质量序列读段，也有许多其他的衡量指标。包括序列读段在不同链之间的平衡（链偏倚）和等位基因变异在不同链之间的平衡（等位基因百分比）。忽略系统误差，等位基因的数量是随机的和符合泊松分布的，因此可以估算假阳性和阴性误差。例如，当一个样品中少于 30%的变异读段是在 30

倍的覆盖度下观测到的，那么该个体是杂合子的概率小于 2%。很多软件包通过不同的方式使用这些指标来识别实际的变异。

总之，虽然 Sanger 测序和 NGS 一开始采用了相似的度量指标，Phred 值和 Q 值，后者增加了许多其它有用的指标来评估总体质量，不只是在碱基识别阶段，还包括了最终获取变异和频率的每一个步骤。这些指标可用来评估单个碱基的质量和整个序列读段的质量是否符合可接受的分值标准。在接下来的拼接过程中就可以自动过滤碱基和读段了。通过这些分值，可以应用一个过滤仅保留符合质量值，比如 Q30 的读段用于后续分析。此外，衡量相邻碱基的质量值是非常重要的；大的偏差可能表示一个有待深入研究的问题。

### 9.4.2 数据清理：生成准确的序列文件

#### 9.4.2.1 文件格式

不同 NGS 平台的原始数据文件目前尚没有一致的或者标准的质量指标；但是，一个标准的文件格式正在逐渐得到一致认可，即 FASTQ 格式文件。FASTQ 文件是一个特定的文本文件格式，用于保存生物序列（通常是核苷酸序列）和相应的质量值。尽管这一格式正在成为标准，但是不同平台产生的内容和质量值的多样性，使得直接比较不同平台产生的数据是很困难的。除了 FASTQ 文件中的碱基识别，还有每个测序反应过程中产生的多个文件和文件类型，并且每个都包含不同的值，质量值，以及过滤低质量读段以便下游分析的特异的仪器运行指标。临床实验室需要花时间分析验证所有的文件类型并确定要保留和要丢弃的文件。处理和保存文件的评估基础取决于临床实验室的分析预期。使问题复杂化的是 NGS 技术由商业供应者开展，并且化学试剂，硬件，软件都定期更新。对临床实验室来说这些更新是破坏性的，在大多情况下，需要额外的分析验证。建议保持与商业供应商的持续联系。

#### 9.4.2.2 序列评价

NGS 反应过程中原始数据的数量和规模巨大，使得其保留、存储和检验都很难而且价格不菲。在临床检测的验证过程中，需要执行一个稳健的生物信息学分析流程，低质量数据（由质量值决定）在数据分析之前自动过滤、以避免假阳

性。一次 NGS 反应产生的原始数据平均为万亿字节级别，这使得临床实验室必须依赖特定 NGS 仪器内置服务器提供的碱基识别（base calling）算法。

**10. 原始测序结果的比对，拼接和评价**

经典的 Sanger 测序和 NGS 测序技术之间是有差别的。Sanger 测序在 1000 bp 及以下的范围中已经、并继续发挥重要价值。NGS 测序则允许以更划算的方式测定基因组更大的区域。两者分析过程的基本差别如下图所示：

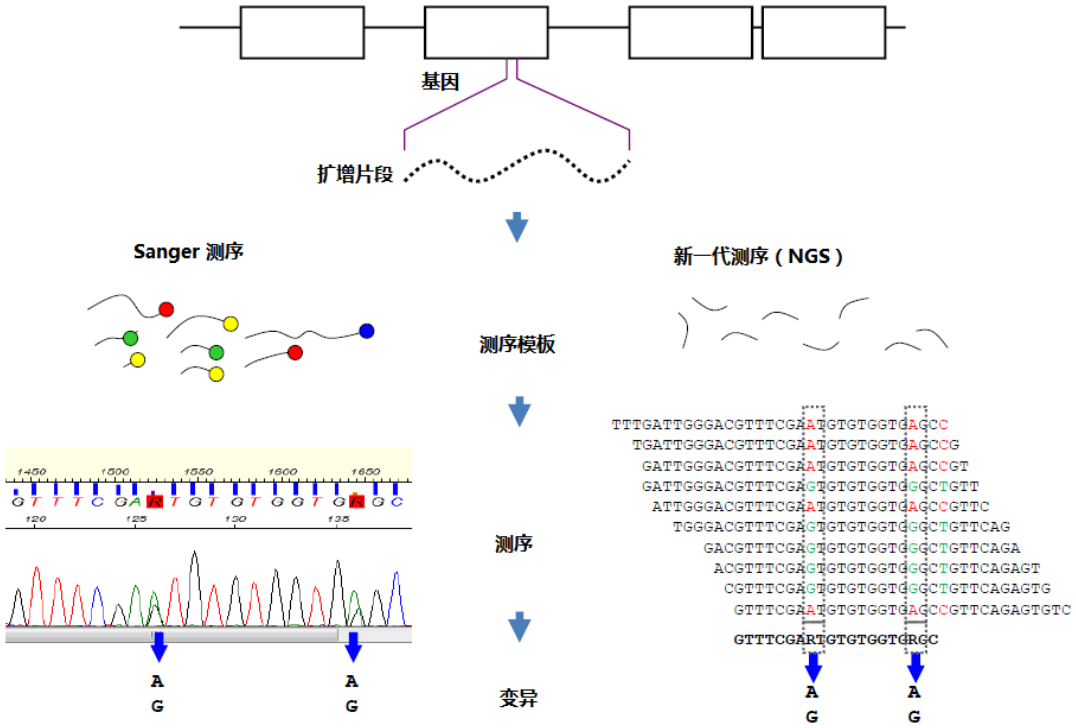


图 2 Sanger 测序与 NGS 程序策略的比较

图 2 中对 Sanger 测序（临床应用中通常采用双向测序）得到的标准电泳图谱与 NGS 测序的结果进行了比较。前者标准化的电泳图谱描绘的信号质量和强度与噪点有关。在杂合位点，两个等位基因的信号大致相同，但通常低于纯合位点。由于信号混合，杂合位点的识别更具挑战性。连续的未知亚群的变异受到测序反应中光学检测能力限制。在 NGS 测序中，每个等位基因是独立测定的；因此，没有信号混杂，能更容易的检测亚群。然而，检测变异时必须考虑总数和采样深度的统计学局限。两种方法都包括需要量化每个碱基的质量值，与参考序列的比对，和一致的碱基识别等问题。

## 10.1 Sanger 测序

### 10.1.1 序列比对

对于来自 Sanger 测序和其他长读长的测序技术的 DNA 序列，可以使用的开源的和商业的软件来进行比对；在某些情况下，能进行图形化检查。序列比对的正式方法采用全局或局部优化。全局优化根据整个序列区域尝试找出最佳匹配。局部优化在两两成对的序列中识别相似区域，采用平铺式（tiled approach）完成序列比对或拼接。对 Sanger 测序的比对和变异识别都有常用的软件工具。常用经典的比对算法，比如 Needleman–Wunsch 算法和它衍生的算法，用于全局比对优化。Smith–Waterman 算法及其衍生算法用于局部比对优化。序列比对过程中混杂的因素还包括单核苷酸多态性，插入缺失，重复长度的差异和结构重排的存在。

### 10.1.2 序列评价（Sequence Review）

#### 10.1.2.1 软件选择

因为 Sanger 测序是一个有广泛基础的成熟技术，有许多高质量商业软件包可用于序列数据的展示和检验。这些软件可以从仪器制造商和第三方获得。它们最显著的差异在于可用性功能（例如，数据显示，操作的易用性和类型，文件存储和检索，以及自动处理）。值得注意一个差异在于识别混合碱基所用的算法。其对高质量的，均衡的数据，这通常不是一个问题，但对于低质量的数据区域会有不同的结果。也存在为特定应用而设计的程序包（例如，HLA 分型或病毒抗性分型），增强了其与应用直接相关的功能。这里建议临检实验室确认一个软件包提供的功能符合实验室的工作流程，并且应用于临床检验结果分析的软件设置要进行性能验证。

#### 10.1.2.2 Sanger 测序数据软件的相关解释

从 Sanger 测序分析获得的原始数据代表了核酸片段在凝胶基质中通过的迁移性差异。片段的迁移性取决于它们的大小和在检测过程中使用的特定染料。分析软件使用迁移变动文件，弥补了由于染料存在引起的迁移变化。使用与燃料对应的正确的迁移变动文件才能正确的解读结果。

进行诊断序列分析的实验室应当证明测定序列的质量。为此目的许多的软件包提供“质量值”选项。在计算“质量值”时许多因素要包括进来，可能考虑到：

- 信号强度：指峰的高度，并可能以相对值或者与标准（电子或对照样品）序列的标准化来表示
- 信噪比：同一区域实际信号（峰值）与其他背景信号（峰值）的比值
- 重叠：在某种程度上相邻的峰占据相同的或不同的位置
- 信号强度的损失：在何种程度上信号强度随着序列的分析过程递减
- 基线的变化：持续信号强度下基线的维持程度

  –信号强度可以是绝对值或与某些标准进行正太化后的相对值（电子或对照样品）。
- 压缩：与后续的峰的相对间距的变化

尽管质量值可以有效的证明测序的总体质量，但这一综合分值不一定能够反映特定的问题。为此，电泳的操作员认证是保证整个流程质量的关键。得到的序列应经过直观评估，以确定上述各项是否会不利于数据的解读。

#### 10.1.2.3 原始数据解读中可能存在的问题

对操作人员来讲，需要能够理解和辨别测序结果在临床上是否合理、还是有可能仅仅是由于技术平台或分析软件所致的问题，这对于结果解读至关重要。问题及其产生的原因，往往可以通过原始或改进的迁移变动模式，电泳图谱，碱基识别，和质量值以及软件提供的其他指标进行推断。同时需要注意，当重复样品的反义链也进行了测序时两链是否一致，尽管其中一条链周围序列的组成可能比另一条链包含更多的信息。参照组的序列质量信息在用于比较以发现问题时也很有用。

低质量数据可能来自多个原因。其表现为：

- 没有可识别的信号
- 噪点导致丢失或错误的碱基识别
- 碱基识别开始之后信号丢失后

- 意外停止

有时，软件解读的序列数据似乎显示为混杂的，但事实上是纯合序列。这种情况可能是由于测序反应中引物的数量大于 1，存在多个引物结合位点，或者模板存在二级结构导致扩增过程中跳过。

核酸测序仪收集数据过程中工作参数的改变可能导致异常的结果。测序反应过程中采用标准控制以保证技术平台的运行符合预期的。

低质量测序反应的故障排除是复杂的，并且可能涉及检查多种可能的原因。如果不能解决问题，操作者应咨询制造商或者测序反应试剂盒销售商、或者技术平台来帮助确定和解决困难。

#### 10.1.2.4 关于测序结果的手工编辑

在某些情况下，在拼接和比对处理之后可以允许在重叠序列的软件无法识别的特定位置进行图形化比较。与软件解读不一致的序列编辑应进行记录并证明。记录应该包括在文件应数据，数据的检测过程，并使决定编辑序列的明确解释。通常，软件包会保留原始序列以便重新审查。最好是对序列编辑让第二个熟悉测序方法的人进行复查（建议采用匿名的形式进行核查）。

### 10.1.3 混合核酸序列的检测

进行测序的样品可能混合有其他核酸。这可能不是预期的，但是必须通过软件和操作人员的图形化检查对其存在进行评估。

存在混合核酸序列的可能原因包括：

- 杂合度
- 混合种群（例如，传染病，肿瘤，非肿瘤细胞）
- 嵌合体□
- 扩增同源区域造成的片段
- 与对照材料反应
- 其他多重扩增产物（有意或无意）
- 污染

如果测序方法本身就是设计用于检测混合序列的，例如检测野生型背景中的体细胞变异或者病原体中的耐药菌株亚群，检测混合位点的能力至关重要。第一步是参考样品或材料的获得，可用于生成已知序列的混合样品，例如序列 A 或序列 B。这些样品必须经过完整的方法流程，包括样品制备和使用用于病例的分析方法进行分析。必须分析足够的样品和混合样品的重复以确定混合水平（例如，30%的序列 A 与 70%的序列 B），以保证可重复性检测在 95%的置信区间内。通过 Sanger 测序检测混合度通常限制在 15%至 30%，某些应用中可达到 10%。低杂合水平的检测可以通过单克隆测序或等位基因特异扩增产生；然而，通常情况克隆对临床实验室是不切实际的并且有污染风险。NGS 技术大幅度提高了分析灵敏度，是需求高灵敏度分析的理想选择。

同时很重要的是，变异的所有可能来源，如仪器的运行性能，不同批的试剂，扩增和测序方法的变化，和操作者的差异，也都可以在混合研究范围内。

双向测序可以提高混合序列碱基识别的可信度；然而，如果需要最高的灵敏度，则可能在一个方向上明显而在另一个方向不确定。造成这种情况的原因可能是，DNA 测序酶将不同的 ddNTPs 加到链上，因此在一个方向上的峰值会很强但在相反的方向上就没有这么强。另一种可能性是，序列不同方向的背景干扰是不同的，其中一链检测值大于干扰，但另一链没有。

如果尽可能高的灵敏度是重要的，那么可能的话应该验证双向的数据分析。置信水平应为单、双向测序分别设定。对于临床 Sanger 测序双向测序是一个标准的最佳做法，病例的任何变异都可以得到检测和验证。

## 10.2 NGS 测序

由于 NGS 获取序列信息的方式，必须采用含多种分析工具的生物信息学流程来进行序列变异的检测。生物信息学分析流程通常包括序列比对或拼接，变异识别，功能注释，临床相关的序列变异鉴定。虽然已经有相当多的成功应对这些新兴的生物信息学的需求，这仍然是一个活跃的发展区。

### 10.2.1 序列变异评价和识别的一般原则

由于 NGS 产生很大的数据和很短的读段，高通量数据的评价需要更高的计

算能力。有许多不同的 NGS 仪器，并都拥有各自特异的数据集和数据类型。NGS 仪器产出的最常见的数据类型包括核酸序列和对应的反映碱基识别可信度的质量值。这些数据的一种常见格式是 FASTQ 格式。FASTQ 文件可用于比对软件去寻找在大的参考序列上的最佳比对位置。同时如果没有参考序列，FASTQ 文件也可以用于从头拼接。序列比对可以通过商业软件包或多种开源的软件来完成（见附录 B，表 B.1）。每个读段的比对信息，比如特定的染色体，定位位置，匹配的可信度，方向，比对的缺口等，储存在第二个文件。流行的 NGS 比对格式是.SAM 文件，其二进制格式是.BAM 文件。最后，从比对好的序列数据中进行变异检测方法有很多。包括检测 SNVs，结构变异和 CNVs 等变异。

一个重要的考虑是，对应于生成数据的特定技术平台，文件格式是类似的，但内容是特定的。因此，在某些情况下，可能存在互相操作的局限性并且数据也不一定有直接可比性。

#### 10.2.1.1 参考序列的选择和使用

完整的参考序列应该可以从公开的网站，如加利福尼亚大学圣克鲁斯（http://www.genome.ucsc.edu）或美国国立生物技术信息中心（http://www.ncbi.nlm.nih.gov/sites/genome）网站下载。在许多情况下，仪器的制造商提供获取需要的参考序列格式的步骤指导。推荐使用最新的参考序列，虽然有时需要根据最新的参考序列对分析流程重新进行确认。下载的参考序列应本地存储归档，因为下游分析依赖于参考序列中的精确序列、注释和染色体排列顺序。应注意确保参考序列不被改变或修饰。除了下载参考序列的名称和版本号（如 hg18/build 36.1 或者 hg19/build 37）之外，在序列分析中用到的任何参考序列的确切名称和本地文件路径都应在结果报告中如实记录和引用。

在选择参考序列时的一个重要考虑，是确定使用全基因组序列还是靶向目标区域序列。一般来说，针对目标区域的测序包括构建全基因组文库（不管采取何种靶向富集方法），都会包含大量“脱靶（off-target）”读段，因此需使用整个基因组进行读段比对。这将减少来自不同基因组位置、但是序列相似的读段的错配，以避免变异识别时出现假阳性。比对到多个不同区域的读段将被舍弃，而且具有较低的比对分值。它可能只使用目标区序列读取映射方法，如 PCR，扩增

特异性高的目标区域和产生可以忽略不计的目标序列。只使用目标区域进行读段比对在 PCR 等方法时是可行的，因为它能高度特异的扩增目标区域而只产生微不足道的脱靶序列。然而，如果靶向参考序列被使用，脱靶量的定位仍需验证。

假基因是 NGS 面临的特别挑战，因为 NGS 一般读段长度短，导致许多读段的比对结果不唯一。大多数假基因所致的问题可以通过仔细评估和改进富集策略来解决。例如，使用 PCR 富集时，引物应锚定在特异的区域以确保只扩增目标区域而不包括假基因区域。假基因可以在样品的定位/比对过程中通过移除定位到基因组多个位置的读段来识别基因组。然而，去掉读段可能导致在目标区域的覆盖缺口。假基因导致的问题还可以通过使用更长的读段、双末端（paired-end）读段以及配对（mate-pair）文库构建技术等来解决。

### 10.2.2 拼接和比对

#### 10.2.2.1 算法

序列比对算法和变异识别通常是商业仪器软件的一部分，有时候这些都是开源软件包的修改版。有许多开源和商业比对算法，针对具体应用的算法适用性需经过确认。NGS 技术产生的读段数目是很大的（通常是百万级别），但是读段的长度比 Sanger 测序短。挑战在于在合理的时间内将 NGS 的读段进行比对。NGS 文库的构建可以借鉴 Sanger 测序即测定插入片段两端的读段（双末端测序）。这种测序区域的制备可跨过大的插入和缺失的结构，并允许通过小的测序读段构建大的叠连群。从头组装对于研究区域是未知的没有参考序列时是非常重要的。

大并且复杂的重复结构对测序错误特别敏感。要重建复杂的结构，可能需要建立一个多态性的大叠连群。这是通过使用能区分该结构内不同区域的多态性来完成的。其关键是识别高质量的多态性以用于拼接重复结构，而那些低质量的应排除在拼接之外。

NGS 分析可用于靶向的（对于选定的区域）或者非靶向的基因组研究。如果存在复杂序列区（如结构变异区）更应注意保证拼接装配质量。复杂序列区包括 CNVs，大的插入缺失，原位倒置，大型复杂的重复，易位等。它们每种都有需要通过拼接过程分析和质量控制解决的问题。并且都会在很大程度上依赖于使

用双末端读段以多种方式来测通或跨过该区域。例如，不一致的成对的双末端读段，其中一条定位到染色体预期插入片段大小位置之外，则可能预示着大的缺失或易位。易位可通过检查不一致的双末端读段，其中一端定位到感兴趣的区域而另一端定位到不同的染色体上或者同一染色体的不相邻的区域（由于大的缺失或倒置）。使用不一致的双末端读段识别易位时应小心，因为重复区域会导致假阳性的易位事件。在这种背景下，一条读段定位到感兴趣的基因，但成对的另外一条读段定位到相邻的重复区域，物理上接近前述读段，但在基因组中多次出现，所以表现为易位。

未知或者假定为未知的区域越多，拼接组装过程就会越复杂、越耗费时间。例如，没有参考序列供定位的从头组装，需要的计算资源远远超出有参考序列的定位组装。在对新物种（比如细菌）或者人类基因组中发生大范围结构变异的区域进行测序时，从头拼接是非常有用的。

#### 10.2.2.2 软件选择

实验室应考虑他们的特定需求和资源选择分析软件。现有许多种开源的公开可用软件和商业软件可用于 NGS 数据分析。大多数制造商会提供与仪器配套的完整分析套件，这些程序可用性较高因为他们针对仪器进行了优化。制造商也适应开源软件和格式，使其更容易整合可替代的相关开源软件。开源软件比大多数商业软件包可提供更高的定制程度；然而，这些开源软件的实现可能需要实验室现有条件之外的专业的信息技术和计算资源。另外，许多软件包可能需要扫描 NGS 的变异图谱，特别是结构的变化，包括插入，缺失，和易位。然而，软件的新的分支或可替代软件包，需要更多的生物信息学支持，质量控制和维护。

现有大量的公共软件包，可以替代或增强制造商软件算法的结果（见附录 B，表 B.1）。这里提供了部分程序清单，但这个名单在不断扩大，希望深入研究生物信息学分析的用户应该查阅最新文献，联系开发者，并为包括提升特定临床应用在内的复杂提升做好准备。根据详细的 RNA-seq 流程，使不同的组件协同工作是很正常的。一个典型的情况可能是，制造商提供的变异检测算法存在一些不足，但不是没有。例如，对 CNV 或微卫星的发现和基因分型可能是不充分的，需要一个额外的算法来增强数据的结果。

选择一个数据分析软件包时，需要考虑的一般要点包括：

- 文档和程序是否有很好的维护？
- 输入/输出数据类型的是否常见？
- 程序是否使用可接受的数据存储和传输的方法？
- 程序是否提供了一个测试集，可以用于校准输出？
- 程序和文档示例与预期使用情况是否密切相关？
- 程序安装和连接其他程序的难度如何？
- 程序是否依赖于其他需要集成到以流程中的组件，这些组件与其他组件是否存在冲突？
- 流程中每个安装的程序，或者需要取代的程序，通常有许多数量和使用示例需要进行评估；然而，应该认识到整个流程也需要重新评价。
- 是否存在该程序与同类程序中其他程序比较的评价或校准？

对于后者，通常有多个权衡进行比较，不幸的是，许多评论或直接的对比只是证明了速度或某种程度的假阳性和假阴性的检测。然而，上述所有的项目都是重要的，没有哪一项更为重要。权衡通常依赖于那些“互相协作（play well together）”的整个流程。

不同选择的影响还包括整个流程验证的必要性。即使只有一个因素的变化，也需要对其所有的使用进行记录和验证。

NGS 数据的比对和分析需要大量的计算基础设施。可以与已知的参考序列进行比对；或者在无参考序列情况下，将读段连接在一起形成大的叠连群，即从头测序。当存在已知的高质量参考序列式，一般应采用与参考序列比对的方法，因为从头方法计算更复杂并有更高错误率。NGS 产生的文件都比较大。例如，一个未压缩的 30 倍覆盖的全基因组测序结果约有 500 GB 的数据。200 倍覆盖的外显子测序通常需要 50 GB 的磁盘空间。这种大文件的分析不仅需要足够的磁盘空间，而且要求更多的计算机内存（RAM）。用户在评估 NGS 的计算需求时应意识到这些问题。

#### 10.2.2.3 数据的改善和重新比对

当对较大的区域进行测序时，特别是在较低的覆盖深度下，有必要对数据进行重新比对。例如，测序数据中的小的插入缺失可能由于初次比对的复杂性会被误解为 SNVs。这些错误可以在比对之后的再次比对提供给软件的周围已知的插入缺失进行部分纠正。其次，碱基质量值可能需要重新校准，因为机器得到的值可能是非线性的而且倾向于高估数据质量，特别是在高质量值区域。质量校正通常考虑碱基在一个给定的读段中的位置，该碱基前面的序列以及测序仪的误差。

如果序列来自于探针介导捕获方法或全基因组文库，那么应该实施额外的质量控制步骤。对每个位点进行高测序深度的检测是有必要的，相同的读段并是不独立的观测（采样样本）。重复的读段，定义为双末端测序读段具有相同的起始位置，彼此完全重叠，代表了可能存在的偏差应该在数据的重新比对过程中删除。通常来讲，与双末端读段相比，从单末端读段数据中更难识别重复读段。重复读段一般来自于文库扩增的过程并因引入变异等位基因频率偏差而导致问题。在比对后的清理过程中，数据文件中的重复读段被标记，以便在后续的变异识别过程中舍弃低质量的重复读段。重复读段的比例应该列为质量控制的一部分。重复读段比例增加可能表明了文库的低复杂度，并且检测低频等位基因变异的能力降低。

### 10.2.3 序列变异识别

#### 10.2.3.1 一般原则

作为质量控制的一部分，平均/预期的覆盖范围和每个碱基的覆盖深度都应进行计算和在结果中报告。低覆盖深度区域应注意并在随后的分析中排除。在一般情况下，覆盖率低的区域往往含有较高 G C 碱基百分比，导致较高的 Tm 值和文库构建过程中效率较低的扩增；这一发现是在富含 GC 的第一外显子尤为常见。在全基因组和 PCR 扩增富集的测序数据中覆盖深度趋于相对的一致，而覆盖深度往往是相对均匀的丰富的序列数据，但在基于捕获的富集方法中是高度可变的。当采用了任何富集策略时，对于识别变异来说找出所有覆盖深度不够的位点特别重要。低覆盖度的许多位点可能在临床上是重要的，缺乏获取可靠变异识别所需的覆盖度应该予以报道。从覆盖深度不足的区域识别变异必须与有足够覆

盖深度但未检测到变异的区域清楚区分开来。

当一个测序反应的序列进行了识别和比对之后，应对所有的变异进行一致的识别步骤。读段的质量值由测序仪的软件生成，尽管不同的 NGS 系统之间的细节不同。因为每一个读段是独立测序的，一致的流程对于同一区域同一样本来说将代表多个、独立的抽样事件。因此，对初始碱基质量识别和抽样要求的数量（覆盖深度）使用合适的阈值，可以确保在每一个识别的位点具有非常高的精度。

使用 NGS，不同类型的序列变异可以在二倍体或非二倍体/群体中等应用进行检测。

除了在生殖细胞或体细胞基因组发生的 SNVs，结构的变异，包括小的插入缺失以及较大规模的，都可以通过采用的比对程序来识别并且完全依赖于比对。不同种类不同大小的插入缺失需要使用不同的比对算法来获取最好检测结果。小的插入缺失变异的识别算法一般整合的 SNV 的识别软件中。更复杂更大的变异则需要不同的算法。多种不同的分析都需要对其性能进行验证，并且对于要检测的每种类型的结构变异/插入缺失都需要建立各自恰当的标准和 LoDs。

基于序列分析检测结构变异的方法一般有四种。理论上，读段对，读段分解和拼接的方法都可以用来检测所有类型的结构变异，但是由于变异的序列特征和读段序列的数据特点各自表现出不同的偏差。然而，基于读段深度的方法可以用来检测剂量改变的变异，如缺失和重复，但不能区分不随数量改变的变异，如易位和倒置。简要来说，读段对方法通过双末端读段的定位信息与预期插入片段大小的不一致性以及链特征进行分析。敏感性，特异性和断点的准确性，依赖于读段长度，插入片段大小，和物理覆盖度。读段覆盖深度分析通过序列覆盖度的增加和减少来相应的检测重复和缺失，并能预测的基因组片段的绝对拷贝数。读段分解算法能够通过读段与参考序列的比对来检测所有变异类型的准确断点；然而，它们通常需要比其他方法更长的读段并且在重复和复制区域表现不佳。拼接算法在所有类别结构变异的断点检测分辨率最高，但对短序列和插入的拼接组装往往导致在重复序列和复制区的叠连群/大叠连群。

这四种通过序列数据检测结构变异的方法都不是综合性的。当许多算法和实验方法应用到相同的 DNA 样品，有很大一部分变异仍然是某个特有的。每种

方法都有不同的长处和短处，取决于变异类型或结构变异位置潜在的序列特征。虽然读段深度是一种准确地预测绝对拷贝数的常用的方法，覆盖度的标准化这一关键步骤是很困难的，依赖于靶基因的数量和它们在基因组的位置（比如，所有的都在少数几个染色体与分散在许多染色体）。读段对的方法是寻找易位，倒置和大的缺失的强大工具，但难以解决重复区域模糊定位，并可能会导致较高的假阳性检出率。

#### 10.2.3.2 单核苷酸变异（SNV）

如图 3 所示，NGS 的独立采样方面也提出了挑战。二倍体基因组中检测出变异的概率是依赖于独立的采样数。具体而言，随着覆盖深度的增加，在杂合的位置检测到变异的概率增加。随着采样次数的事件的增加，任何位点误检的可能性（假阳性）在任何位置也增加；因此，任何 SNV 检测工具的灵敏度和特异性都需要实验室进行仔细的验证，从而为特定的分析流程确立适当的阈值。

首先，所有的碱基应基于初始质量值进行过滤；质量值大于 20 或 30 的读段对于具体的应用非常有效。当识别变异时，超过一次的检测到第二个等位基因是很关键的。下图显示了一个二倍体的例子；采样 12 次，得到第二个等位基因（如果存在）至少一次的观测率 99%。对杂合位点的准确识别，对所有要报告位点的平均覆盖深度至少应为 20 至 30 倍。由于检测受多种因素的影响，检测特定平台特定序列区域的特定类型变异所需要的具体范围应该进行验证。

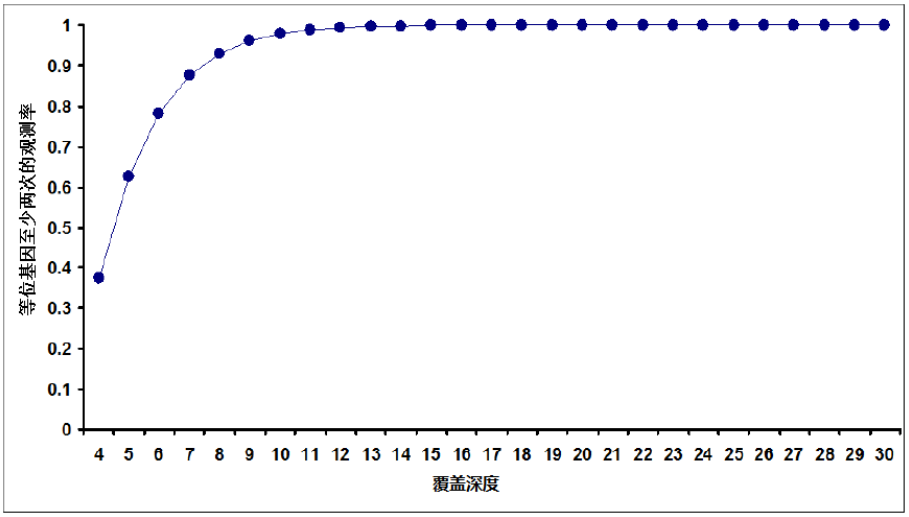


图 3 覆盖深度与特定变异的观测率之间的效应关系

由于被检等位基因变异率不同，特定区域的扩增也需要进行大的改变，而个别样品低频次要等位基因表达。对 NGS 过程中这些区域的表现进行提前了解并在验证过程进行确定将提高检测效率。对于任何意想之外的预期偏差也应该建立和监测其表现值。如果该区域使用了靶向方法的选择，由于引物和扩增的偏差应采取额外注意。很重要的事，在所有过程中需要更深的覆盖度导致的错误都需要在测序方案的优化中来决定。

确定变异存在的关键因素包括：

- 初始碱基识别质量
- 进行识别所需的采样数（覆盖深度）
- 相对于第一等位基因，第二等位基因检测到的次数
- 包含变异的读段其比对质量值
- 检测到变异的读段其正向与反向的数量（方向偏离）

使用变异识别软件时，上述参数一般纳入进来并进行评估。NGS 中的独立取样和足够的测序覆盖深度使之能够检测频率在 50%以下的变异（50%是杂合变异的概率）。需要警惕的是杂合区域由于处理过程的原因也可能与 50：50 的比例差别很大，验证过程需要建立其表现标准（例如，需要设置不同的稀释倍数来验证感兴趣区域的变异）。非二倍体的变异检测包括在传染病样品中识别变异其目的是查找潜在的耐药亚群，在肿瘤样品中检测变异可以发现哪些变异对化疗决定有重要影响或者肿瘤细胞在哪些周围基质中稀释，以及检测异质性。在每一种情况下，都可能发现变异，但要使检测可信，需要彻底的调查和并为不同的目的建立特定的标准（例如，检测样品中频率为 5%的变异）。上述分析原则不会改变；然而，特定的阈值，分析需求，以及报告的特异性和敏感性可能是完全不同的。

#### 10.2.3.3 插入/缺失（Ins/Del）

NGS 可以识别插入缺失所致的结构变异，长度范围从 1 个 bp 长度到几 Mbp 不等。应该指出的是，插入缺失的检测容易产生假阳性原因在于初始比对误差，测序过程的滞后误差，以及人类基因组中的重复区域。存在多种检测插入缺失的软件，他们分别有自身的局限性。小的插入/缺失（小于高通量读段长度的 15%）通常可以使用一般的变异检测工具，依赖于含插入缺失读段的初始的正确定位比

对。中等大小的插入，其插入片段的大小接近于读段长度，其检测是具有挑战性的，因为含插入片段的读段可能与参考序列相似度不足以进行合适的比对。这类中等大小的插入可以使用专门的软件来检测，它们对未定位的读段进行类似从头比对的方法比对到本来部分匹配、或有可剪接结构的区域。大的插入缺失（> 500 bp）可以通过成对定位的双末端读段评估插入片段大小进行检测；过大的插入片段可能表明插入而较短的插入片段可能表明缺失。

插入缺失的检测存在困难，建议各实验室严格评估比对软件包，使用已知的控制条件即插入缺失的大小与临床评估一致。如果大范围的插入缺失需要检测，则可能需要多个检测工具。

#### 10.2.3.4 拷贝数变异（CNV）

CNV 分析也可以通过比对大量的 NGS 读段来进行。大部分通过高通量数据检测 CNV 的方法依赖于读段在参考基因组的定位并找出覆盖度不一致的区域，即表示一，三，四，或更多的拷贝数。所有的方法需要初始覆盖度与无关标准对照进行标准化的步骤，配对的正常组织（癌症研究），或者样品总体覆盖率。覆盖度标准化可能会受到目标区域基因的数量和组成的影响，因为基因组的某些区域更容易发生拷贝变异，特别是癌症中。在 GC 核苷酸丰富的区域，或者有低复杂度的序列，往往具有数量较少的读段。这些时候必须小心，因为这些偏差会影响拷贝数的计算。大多数 CNV 类型理论上可以识别但是他们与建库过程中插入片段的大小是敏感相关的。各种方法都没有明确的优势；可能需要多算法来进行全面的分析和评估。针对这个过程已有评论文章发表。良好的覆盖度均一性和“正常”的知识库是进行可靠的 CNV 检测所必需的。

#### 10.2.3.5 易位（Translocations）

在全基因组和靶向 NGS 数据中都可以通过专用的软件在 DNA 或 RNA 水平检测易位。总的来说，这些软件通过寻找双末端读段的在不同染色体的不一致比对来检测易位。双末端比对的方法会导致假阳性增高，因为基因组中存在许多重复序列并可能导致读段的多重定位。最近的易位检测方法通过跨越易位边界的嵌合读段或识别大量可剪接读段，可以降低检测的假阳性。

### 10.2.4 群体测序的具体问题：肿瘤和传染性疾病

群体测序是检测混合群体中亚群的变异，其中包括肿瘤相关和 HIV 等传染性病原体。序列的覆盖度水平和实现方式取决于具体问题。例如，问题可能是在某个特定的灵敏度下群体内是否存在某个频率范围的变异。通常，更高的测序覆盖深度可以达到所需的检测能力和灵敏度；然而，成本也是一个因素。在肿瘤诊断中的应用，样品通常是肿瘤和非恶性细胞混合的异质性克隆。正常组织和肿瘤样品一般都进行测序和比较以确定变异是生殖细胞还是体细胞来源。计算变异识别灵敏度和特异性的方法之一是使用一系列稀释的包含已知变异的 DNA（阳性质控品）来模拟等位基因频率的下降。因此在不同的覆盖深度检测特定频率序列变异的能力可以通过一系列的预实验建立。建立个体碱基水平的错误结构可有助于敏感性。最后，应注意对于所有检测到的变异保持低的假阳性和较高的阳性预测值。

群体测序的第二个例子是传染性疾病的特异病原体检测，例如 HIV 病毒的趋化性。在这种情况下，检测群体多态性的能力可以通过主样品的多等分来得到更好的低水平多态的随机概率。关于 HIV 病毒等研究中确保深度测序中的准确采样，已经有许多技术文献发表。

### 10.2.5 计算问题和解决方案

NGS 技术产生很大的数据集（通常是 10 GB-100 GB 之间），也因此产生了下游生物信息学分析中的一系列问题。仪器相关的基础设施是复杂的，尽管制造商正在努力克服许多的 NGS 运行和分析的复杂性，整个工作流程中仍有许多空白需要错综复杂的生物信息学知识。

NGS 生成了包括测序反应图像（大小往往几百万兆字节）在内的非常大的原数据集，这些一般不需要归档。包含原始序列识别和相关质量值的中间文件一般为 SCARF 或者 FASTQ 格式。然后这些文件与参考基因组进行比对得到比对数据文件（一般为方便阅读的.SAM 格式或压缩的二进制.BAM 格式）。如果必要，这些比对文件可以存储而中间 FASTQ 或者 SCARF 文件可以再生。在某些情况下，在从原始的 FASTQ 或者 SCARF 文件生成.SAM 或者.BAM 文件时数据丢失，可能导致所有数据都无法恢复。比对数据文件时，通过下游变异检测或变异识别

软件用于测序数据的序列变异识别。这一步通常生成一个变异识别格式的文件。序列变异的识别可能出现很大程度的差异，取决于所使用的分析程序和能够检测的变异的范围（例如，SNP，插入缺失和易位）。因为分析技术仍在迅速发展，变异识别文件应该不是唯一的测序存储文件。建议将.SAM 和.BAM 文件保存，以备未来数据分析所需。尽管已经进行了压缩，.BAM 文件仍然很大，经常超过 100 GB 大小。关于存储哪些文件和保存的后面将有详细讨论。

分析 NGS 数据所需的计算基础设施需求非常高。硬件设施通常是由 NGS 设备制造商建议，但是强烈建议进行设备的安装程序之前咨询信息技术和生物信息学专家。从比对到变异识别的分析过程所需要的时间，与可用的计算机中央处理器（CPU）数量有关。

处理这些大型数据集和相关分析的另一种可能性，就是使用云计算或网格计算结构。许多制造商和数据分析服务商正在积极推动云计算设施来帮助小型实验室，并编写大量适用于这些平台的拼接程序。云计算与传统的用户维护的数据中心相比有其优势和劣势。优点在于高水平的可扩展性和较低的入门价格；缺点则包括需要时间将大量数据文件从本地服务器上传到云端，算法是否可以处理大量的 CPU，云端共享存储和 CPU 的情形下的安全问题，以及云计算整体监督控制缺乏和病人的隐私暴露问题等。

比对过程可用的算法越来越多，这确实也是必要的，因为 NGS 技术本身也在不断发展和提高，比如读段越来越长、数据质量也越来越高。虽然有些第三方的比对算法在某些特定情况下可能表现更好，但推荐处于起步阶段的 NGS 实验室还是应该优先使用测序仪制造商提供的算法，因为它可能针对仪器进行了特殊优化。

目前在处理 NGS 数据的软件有许多，从原始碱基识别和读段生成，经过比对，到变异识别。对于用于个体化医学检测的数据分析软件，所采用的特定算法都需要经过验证方可使用。NGS 仪器和算法的整个流程在与临床分析同样的情况进行测试是很关键的（包括程序和程序内的参数设置）。记录和验证整个测试过程非常重要，并且在应用范围内和临床应用的变异类型而言，对已经验证的参数所做的改变，需要详细记录，并再次验证应用范围和各种系统参数，才能被继

续应用于临床。NGS 的数据分析通常包括仪器配套的软件分析和实现最终检测目的的额外算法分析。

### 10.2.6 原始数据解读的问题示例

NGS 分析可能会产生误导性的结果，取决于底层的序列变化和上下游序列的内容。插入缺失检测是一个很好的例子。许多分析程序不会识别超过一定大小的插入变异，超过时将报告假阴性结果。当检测到插入缺失，报告的大小可能是正确的，但准确的基因组的定位可能与 Sanger 测序不一致。如果周围有重复序列或插入/缺失序列定位到多个位置时，这种情况会非常明显。这些错误通常会导致 NGS 报告的插入缺失位置发生改变，结果与 Sanger 测序相比时，导致所报告的蛋白质编码变化不同。引起插入缺失错误定位的因素包括软件和上下游序列，这些都应当在测试验证过程中进行评估。

在高通量数据解读中另一个常见的问题是测序识别错误的低频假阳性。当变异检测能力高于阳性预测值时这也会升高。例如，如果所用测序仪的错误率是 2%，那么这些读段中只有 2%频率的变异将产生明显的假阳性结果，除非分析软件强大到足以检测这类假阳性结果。为了避免低频变异的假阳性，测试实验室应该建立 NGS 的 LOD 和阳性预测值。LOD 和阳性预测值可以通过包含一系列稀释的已知变异 DNA（阳性质控品）来估算。

标签互换，即通过分配给混合样品中的一个条形码或者标签被分配给了另外一个样品，从而会使假阳性变异识别增加。对大多数文库构建和扩增方法，标签互换率较低一般在 0.01%的水平。然而，当互换率高时情况假阳性可能会更高。使得互换率升高的因素可能包括混合捕获方法（如标签的库被一起捕获，而不是单独捕获然后测序才混合），较短的标签序列，标签序列具有较小的 Hamming 距离，以及分解标签时允许标签错配。实验室应针对特定的方法确立互换率来确保不去识别低于互换频率的变异。互换率可以通过与临床样品没有显著的同源性的 DNA 来估计（如，测定人类样品时使用噬菌体 DNA）并测量外源 DNA 与错误标签关联的频率。

## 10.3 单分子测序

单分子测序技术刚刚开始进入科研应用领域。与现有的NGS测序技术相比，这些技术可以免去 DNA 扩增步骤，直接对待测的单个 DNA 分子进行碱基序列的读取。这些测序系统理论上应该比现有 NGS 测序技术更快、还可能获得更长的读段。同时，因为所有 DNA 扩增步骤都会不可避免的引入误差和噪点的干扰，因此，不需事先进行扩增的单分子测序也有可能提高测序的准确性。但到目前为止，单分子测序技术的准确性还没有得到明显突破，而且性价比要远远低于现有的 NGS 测序技术，因此未能实现大规模的推广应用。当然，许多针对 NGS 和 Sanger 测序所讨论的原则，对于单分子测序技术仍然成立，包括读段、拼接和变异识别。建立测序流程 QA 和 QC 的一般原则也适用于这些单分子测序技术。

## 11.质量保证和质量控制

### 11.1 质量保证和质量控制的定义和要求

与个体化医学相关的测序分析过程的质量控制，是个体化医学检测质量保证的核心内容，是个体化临床检验规范化和标准化的首要前提。因此，所有基于 Sanger 测序或 NGS 的临床基因检测项目的研发、试验性能确认/验证以及检验全过程，都需要建立有效的质量保证（QA）和质量控制（QC）体系。

### 11.2 数据评价

测序实验的数据解读，高度依赖于测试的指标、基因组的目标区域和测试结果。测序分析可以用作一种诊断方法，即在个体中查找已知的功能变异。测序分析的主要目的可能是为了证实或排除临床上可能的诊断，从临床上多种可能的情况中进行精细的确认。在这种情况下，变异位点的数据可以从序列的质量、覆盖度、识别的可信度进行检查。因为变异是已知的，对此的识别有助于临床解读。另外，一个人可能会有未知的临床诊断，那么测序分析的目的就是确定最有可能与临床表型相关的变异。所有的测序分析都有可能检测到临床意义和临床有效性不确定的变异，因此，验证试验和适当的后续解释非常重要的。然而，测序分析也可能无法检测到所有的变异，而后者可能是临床上重要的。特别是对未知变异的分析，在验证阶段必须明确变异的类型和范围。数据应当进行与验证的阈值和收集的趋势相关的评价。

### 11.2.1 Sanger 测序数据评价和质量评价

电泳结束后，测序分析软件将电泳数据转换为序列数据。测序反应质量的第一个评估是软件能否为每个样品产生序列。如果在测定大的目标序列时使用了多个引物，比对软件从每个反应生成一致序列的能力是一个很好的质量指标。序列应该可通过自动仪器输出的结果文件来对电泳数据进行交叉引用。

电泳图谱检验的重要特征包括：

- 信号强度（峰高）
- 信噪比
- 净间距或峰的分辨率
- 低背景和噪点（小峰）

碱基间距的规律可能受到富含 GC 的区域影响。相对于分析验证过程中纯合子的峰值大小，背景峰应该只占一个小的比例。当电泳图谱的特征是可以接受的，那么仪器软件通常能够生成高质量的测序数据。通过查看软件识别或者未识别的碱基数量可以很容易的对此进行评估。在评价序列时，应该确定碱基识别的准确范围。如果未识别的碱基数量超过仪器制造商的规格或者实验室通过验证核实研究得到的规格，那么电泳图谱应该经过用眼检测以确定该样品是否应该重新进行反应。对于自动化仪器或商业试剂盒，有关碱基识别准确率的范围应该向制造商咨询。制造商提供的测序仪和序列分析软件手册通常给出了有问题的序列数据和故障排除实例。其他信息可参见网站（http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.html）。

在进行变异识别的序列分析时，检测序列的两条链是最佳的。当不能实现时，需要检查测序分析的质量或者通过重复实验来进行证实。将样品序列的保守区域比对到参考序列是一个有用的质量评估方式。较低水平的一致性表明样品或试验可能存在问题。在许多应用中，预期只在一小部分的序列存在变异，大部分的序列可比对到参考或共有序列。可以通过计算和应用“全碱基准确度（all-base accuracy）”来评估得到的序列的质量。所有碱基的准确度是指，在参考序列的指定区域中，识别的碱基与预期碱基一致的百分比。病人样品和参考序列之间高度保守的序列子集才应被应用于数据分析。测序反应中碱基识别的错误往往分布不

均匀；在序列的开始普遍存在的大量的错误识别。实验室应调整测序反应，通过在目标区域足够远的上游开始测序来应对这一趋势，从而减少对感兴趣的 DNA 的错误识别。

样品序列和参考序列之间的差异可能至少有三个原因：

- 与参考或共有序列相比，样品有预期之外，但真正存在的序列差异。
- PCR 所用的非高保真酶产生的样品与参考序列或者共有序列的差异。
- 自动序列分析软件对电泳图谱的分析不正确，在样品中识别出错误的碱基。

所有碱基预期的准确水平取决于扩增，测序试剂和测序设备的选择。

### 11.2.2 NGS 数据评价和质量评价

进行诊断序列分析的实验室应当能够证明每个测序仪器反应过程中产生的序列的质量。仪器配套的软件提供了有用的运行指标来评估原始数据的质量，除此之外，许多软件包也提供了额外的质量分数。建议临床实验室对每个检测项目均检查这些额外的质量分数和指标。应该以制造商建议的平台的性能指标为指导，来设定确定通过/失败的指标阈值和实验室各自的其他质量评估标准。

NGS 常用的一些质量值和指标包括：

- 忠诚度（Chastity score）：代表测序读段的纯度：本质上是信噪比。低忠诚度分值的读段通常在碱基识别之前被过滤掉。
- 所产生读段的总数量：测序读段在流动槽的密度是固定的。较高的簇密度会降低仪器解读独立读段的能力。当超出设定的读段数目时，背景“噪音”增加，许多读段的忠诚度得分降低并在碱基识别和后续分析之前过滤。
- 通过过滤的读段所占的百分比：即将（通过过滤）得到的读段总数与由于低忠诚度而过滤掉的读段数目进行比较。如果通过过滤的读段所占百分比较低，则表明这是一个次优的测序反应。
- 信号强度：每个碱基的信号强度，特别是在第一个碱基和随后的特定碱基。

- 滞后（Phasing）：每个循环中高于或者低于同时期测序反应总平均值的核苷酸纳入测序读段的比率，导致背景干扰和不正确的碱基识别。

NGS 监测的其他有用指标，包括但不限于以下内容：

- 预期的与观测到的数据产出
- 正向和反向读取的平衡（如果存在）
- 额外碱基的信号强度
- 覆盖度的一致性
- 测序的覆盖深度
- 与互补链的数据比较
- 滞后和超前

注：上述指标在不同的平台可能会有差别。

除上述之外，还会有平台特异和试剂特异的质量评估指标可以使用，建议操作者与 NGS 测序仪的制造商密切联系来建立和验证这些指标。

虽然质量值和仪器运行指标在证明生成序列的总体质量时很有用，但综合得分可能或者不能反映某些问题。在 Sanger 测序，可以通过评价原始数据和处理后的数据来确认整个流程的质量。对序列进行可视化评估来确定是否有不利于原始数据解读的上述因素。

然而 NGS 设备一个测序反应的大数据量限制了操作者评价原始数据和处理后数据的能力。不可避免的是，任何临床检测项目的验证过程中，应该采用强大的生物信息学流程，在分析之前以自动方式过滤低质量数据的测序读段。商业化的可用仪器设备，在不同的文件和格式中提供了多种质量值和指标。考虑到所需的时间和其他资源，建议临床实验室在开始比对之前检查质量值，各项指标和读段总数等情况以确定运行的整体质量。

NGS 结果的质量控制可以通过流程中每个部分的标准来实现，包括特定水平的变异识别的准确度，灵敏度和特异性。尽管不同制造商的高通量系统之间存在不同，许多质量指标的存在，可以帮助操作者决定测序反应的质量是否能够使用以及碱基变异的检测质量能否用于临床相关性解读。NGS 平台指标的详细描

述见前述。需要谨记的是，基于要回答的问题、序列的特点和其他过滤阈值等，数据可以在多个步骤进行过滤并且需要适当调整。

群体测序有特定的指标并可以增加额外的质量控制措施。其中包括计算特定变异的覆盖度水平和质量，以及系统发生树的构建来确认序列符合总体变异和系统发生的规定范围。

### 11.3 序列软件评估和检验

应当在定义的测试范围内对检测变异的软件进行验证，并要按照验证建立的指标和规格来运行软件。任何设置的更新或改变可能会对分析的下游部分产生意想不到的影响，在常规使用前必需经过重新验证。一个地方的改变可能会导致其他指标的阈值发生变化，这也必需重新评估和改善。

### 11.4 NGS 检测实验室的质量管理体系

NGS 与 Sanger 测序虽然都是测序技术，但各自的技术原理和下游的数据分析具有本质的不同，将 NGS 用于临床时也会产生更多的监管问题，因此需要针对 NGS 检测实验室的制定完整的质量管理体系（图 4），以确保 NGS 的检测结果能够用于指导临床决策。

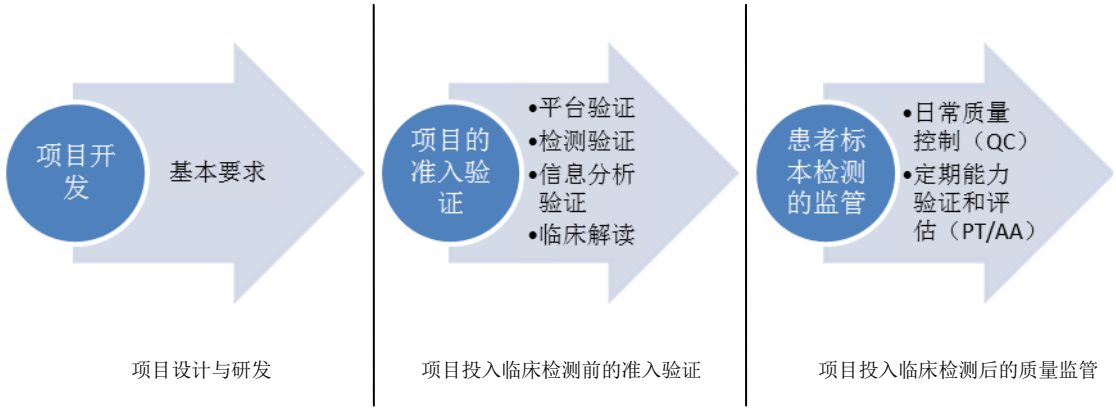


图 4 临床 NGS 检测实验室的项目监管流程

#### 11.4.1 新一代测序平台选择和基本要求

（1）测序平台的选择

在选择购买和配置测序仪器和相应的检测体系时，需要根据检测项目的样品特性和结果要求等，综合考虑以下因素：

A 读段的长度（简称读长）：如果所分析区域含有非常高的同源区域，读段足够长的检测平台才有可能获得准确可靠的数据；而对于判断核酸片断的拷贝数的检测，例如，小 RNA 表达检测、数字表达谱检测、染色体数目异常分析等，仅需要检测较短片断就可以完成。对于某一个特定的测序平台，基于各自的检测原理不同，读长是一个可变范围很小的参数，因此需要根据自身的检测需求选择合适的测序平台。

b 特定区域所需要覆盖的测序深度： 对于突变比率低的样品，例如肿瘤组织的体细胞突变（somatic mutation）、高度异质的线粒体突变等，需要较高的测序覆盖深度。对于突变比率低的样品，验证平台不能采用 Sanger 测序，需要考虑采用重复检测或者其他替代性平台进行验证。

c 测序项目的样品量：不同测序平台对检测样品的起始量要求不同，需要根据检测项目可的样品来源和最大起始模板量进行选择。

d 所需要的报告时间：用于病原体检测、药物治疗指导等检测的项目，都要求报告时间尽可能短，因此需要选择检测流程快速、测序时间较短的测序平台。

e 检测项目的预计检测成本：新一代测序的检测通量大，理论上可以降低单个样品的检测价格，但是具体每个项目运营时的实际价格，还与每次可以上机的样品量有关，因此，除了考虑仪器和试剂成本之外，测序平台的通量、可选择的检测项目等都是影响测序成本的因素。

f 测序技术的总体准确性：对于用于临床基因检测的 NGS 检测技术，要求测序技术的原理和实际验证数据的准确性达到 99%以上，验证方法采用 Sanger 测序技术或其他对应的检测方法。

（2）测序下游的信息分析和测序数据管理

生物信息学是 NGS 下游数据分析的重要支撑平台，它是生物学、数学、计算机三大学科高度交叉融合的产物。为了分析 NGS 产生的海量测序数据，目前已有很多信息分析工具可以完成下游的比对、拼接、变异识别等工作。对于大多数 NGS 平台，需要 NGS 数据分析是从图像文件中进行碱基识别开始，在此过程中，每种平台采用自己开发的算法进行测序数据的质量值（quality score），用质量过滤器（quality filter）筛除质量评分低于标准的片断后，与参考基因组比对（没有参考基因组的测序数据需要进行从头拼接），目前已有许多软件可以完成短读

段的与参考基因组的比对工作。当测序数据被比对、映射（map）到参考基因组序列之后，就可以在碱基识别的基础上进行变异识别，并标注出变异位置、变异等位基因百分率等。以上数据分析流程可以产生一个包含序列变异（SNP、Indel、其他结构变异）、数据总结与丰富的注释信息等的通用格式化文件，称为变异识别格式（variant call format，VCF）或基因组变异格式（genome variation format，GVF）文档。

要完成 NGS 平台下游的信息分析工作，需要配备强有力的信息技术平台，包括强大的数据储存和运算能力的计算设备，并满足空间环境、网络信息存取等要求。另外还需考虑，在检测完成之后，实验室长期保留原始数据是质量管理过程中的重要一环；因为目前对于能否仅保存变异列表文件（如 VCF）还有争议，因为在信息分析流程或参考基因组序列改变时，只有保留原始数据才可能重新进行信息分析。如果 VCF 文件之外，同时保留包含质量信息和比对文件（BAM 格式）的原始基因组序列文件（如 FASTQ 格式）也是一个备选方案。

目前对于下游的信息分析还没有一个标准的、线性化的流程，每一次检测的数据分析过程可以根据检测所关注的数据变异类型进行适当调整，而且可能还需要根据碱基识别获得原始数据的错误来源、测序覆盖倍数等，对分析参数进行灵活调整。此外还需考虑上下游序列可能造成的结果偏倚，主要包括 GC 偏倚和链偏倚。高 GC 含量的基因组区域不易被靶向捕获，捕获后的片断也难以通过 NGS 进行测序。由于很多基因的第一外显子都富含 GC，所以很难获得满意的测序数据。如果大部分的测序数据都仅仅来源于一条 DNA 链，或者变异碱基仅出现在一条链上，就可以判断存在链偏倚。因为错误的碱基识别可能会在一条 DNA 链上累积，来自正义和反义链的测序数据对于确保变异识别的正确率非常重要。在设定信息分析流程时应考虑设置相应的过滤器以整合来源于不同 DNA 链的碱基信息、降低出错率。

对于假基因、同源基因家族等同源性较高的基因区域，很容易因为读段比对时的匹配位置差异，而出现假阳性（FP）或假阴性（FN）结果，尤其在读段较短的 NGS 平台更为明显。因此，在建立数据分析流程时需要包含在这这些同源区域中识别疾病相关变异的针对性策略，例如全局比对之后的重新比对等。

还需要注意的是，人类基因组中的有些复杂区域 GC 含量太高、或为重复区

域，即使加大测序覆盖深度，也不能通过 NGS 技术得到满意的测序结果。而且，现有的人类基因组参考序列仅来源于几个个体，并不能代表全人类的基因组（更有甚者，有一些基因组区域目前还没有参考序列），因此在序列比对和结果解读时需多加小心。

总之，到目前为止，信息分析软件包一直还在不断开发当中（参见附录 B，表 B.1），也没有一个可资使用的参考基因组和分析流程的金标准，因此实验室管理人员有责任确保信息分析流程经过最佳调整优化和验证，而且对于有些基因组区域目前还无法通过 NGS 检测这一事实有正确的理解和认识。

### 11.4.2 检测确认（validation）和技术特征参数

在将检测用于临床工作之前应该对方法进行分析确认和临床确认。检测确认是为了确保实验室自主研发检测项目（laboratory developed tests, LDT）的准确、有效性而进行的分析性能和规格建立过程。在确认过程中，受检实验室必须证明整个测试过程与预期相同，而且能提供可信的结果。

（1） 针对检测平台、检测方法、信息分析流程的确认过程

在应用于临床检测之前，新一代测序实验室必须确保测序技术的分析有效性，并提供相应的确认措施。检测实验室的确认过程包括平台确认、检测方法确认和信息分析流程确认 3 个相互联系的部分。

在接受平台验证之前需建立可用于基因组 DNA、福尔马林固定石蜡包埋（FFPE）标本 DNA、血浆游离 DNA、组织或血浆 RNA、以及病原体核酸等样品处理、测序分析和下游信息分析等整套技术平台与流程，并且机构和空间设置符合原卫生部制定和颁发的医疗机构设置与管理条例、以及核酸扩增实验室管理规范，以确保检测平台提供的测序数据是准确可信的。

（2）特定检测项目确认

需证明特定的检测项目确实可以检出具有临床意义重要的序列变异（或其他类型序列异常）。检测项目需经卫计委批准进入临检目录，或其临床意义应经过国家卫生计生委个体化医学检测专家委员会论证。

（3）信息分析流程确认

确定配置了所需的信息分析软件，可以确保能够有效地分析序列数据，并检

测出变异或其他序列异常。鉴于目前的 NGS 平台种类较多，同一平台的检测策略和技术流程也各不相同，分析软件需要能够根据不同平台和流程的特点调整算法和参数，并有相应流程。

（4）检测结果的确认

对于可能存在问题的复杂基因组片断，需要建立备选的检测平台（例如 Sanger 测序）进行结果确证，以确保测序数据的质量。

（5）检测结果的临床解读

对于检测结果与临床意义之间的关系已经被用于临床的项目，尽可能根据现有的文献结果给出解释，指明检出的变异是否改变蛋白质编码和功能，与疾病表型之间的关系等信息；但是对于检测结果与临床意义之间的关系尚不明确、或者未得到公认的检测项目，需参照第三类医疗器械审批标准完成临床试验，获得能够证实检测结果与临床意义之间存在相关性的实验数据，并经过国家卫生计生委个体化医学检测技术专家委员会论证之后，才能用于临床检测和收费。

以上 5 个验证过程的侧重点各有不同、但又密不可分。在验证过程中要结合实验室开展的特定项目进行，需要考虑检测目标是基因组合、外显子组或者全基因组，以及所分析的序列变异类型是 DNA 突变、RNA 拷贝数变化、染色体数目异常等进行综合考量。

对于每一个实验室进行的 NGS 研发项目，均要求提供详实的技术规范和性能特征，包括检测的准确度、精密度、敏感性、特异性，检测范围，参考范围，以及其他相关的性能参数。除此之外，NGS 实验室还需要设置和监测的 NGS 特有技术参数包括：覆盖深度，即用于特定区域碱基识别的有效读段（reads）的数目，这是确保 NGS 检测结果的准确性、敏感性和特异性的重要参数。对于分析 RNA 转录或者染色体数目异常的检测项目，则需要经过严格测试验证后，明确限定检测所需的读段（reads）总数据量。

由于相对高昂的检测价格以及后期所需的繁复的数据分析过程，限制了采用重复性检测的方法来验证 NGS 检测结果的精密度，特别是需要对大批样品进行检测时更为明显。对于目前的平台验证而言，应该采取最低程度的重复性验证：推荐建立项目所需的检测参数时，采用至少 3 个参考样品，在相同和不同批次间进行检测和比较，以验证平台的精密度。

对于特定检测项目确认，可以将测序深度以及测序覆盖的一致性等参数用于精确性验证，建议参照报告范围和参考范围两个定性参数进行调整。例如，通过限定报告范围可以反映检测项目的检测质量，在参考范围之外的序列变异可能是病理性的，需要采用进一步的策略进行确证。

以上的确认工作的周期要求为：

（1）实验室在正式批准运营之前，必须完成并通过整个实验室的检测确认工作。

（2）获得批准的 NGS 实验室应该每年进行一次确认。

（3）更改实验室环境或检测关键环节时，例如实验室场地更换、更改仪器或关键试剂、检测样品类型变化、原先设定的基因组合中加入新的基因、信息分析软件升级等，实验室必须重新进行性能验证。

### 11.4.3 质量控制（quality control，QC）

质量控制过程用于监控日常 NGS 检测流程的所有环节，包括试剂、样品处理、文库制备、上机测序、仪器设备性能与维护、数据处理（信息分析流程）已经报告生成和临床解读等测序检测前、中、后的各个环节，以确保检测流程的顺利进行和检测结果的准确性。

在日常 QC 工作中，可以采用质控标准品作为内源性或平行对照，以模拟基因组复杂性或与检测目的类似的序列变异类型。

在用于患者样品检测时，需要记录和评价 NGS 性能检测的特征性 QC 参数，包括：覆盖深度、序列覆盖一致性、用于评价碱基识别和比对的质量值（quality scores）、等位基因百分率（allelic read percentage）、链偏倚（strand bias）、GC 偏倚以及信号密度衰减（decline in signal intensity）等，并且与检测验证过程中的参数进行比较。

对于影响临床决策的核酸变异结果，在签发报告之前，需要另外采用备选方法（如 Sanger 测序、SNP 芯片等）对检测结果进行确证。

### 11.4.4 能力验证（proficiency testing, PT）/室间质量评价（EQA）

NGS 检测实验室应定期参与能力验证以及其他外部质量评估项目（如替代性评估 alternate assessment，AA），通过实验室间的检测性能比较，来判断实验

室的检测能力，并发现分析和解释错误，以及 QC、仪器校准、实验设计等方面存在的问题。PT 项目将作为经过认证的 NGS 检测实验室后期复检的重要环节。

PT 的实践操作流程为，提供变异类型已知或未知的标准样品，供 NGS 实验室“盲测”， 室间质控样品应与常规临检标本同时检测，受检实验室将提供检测结果、解释，以及关于所采用检测方法的简要描述。来自不同实验室的数据互相比较以评价室间检测的性能，最后给受检实验室提供总结性评估报告。

对不同受检实验室的能力验证进行比对时，需要考虑不同实验室采用不同技术平台之间的差异。

NGS 检测实验室应该每半年参加一次由国家卫生计生委临床检验中心组织的能力验证，包括检测技术验证（湿实验部分）和数据分析能力验证（干实验部分）。NGS 检测实验室应积极主动参加国家卫生计生委临床检验中心组织的室间质量评价。

### 11.4.5 参考样品（reference materials，RMs）

在上述的实验室检测系统验证、日常质量控制和定期能力验证中，都需要用到参考样品，以监测临床检测项目的质量可靠性。

参考样品应该是均一、稳定的，有可以检测的特定属性（例如存在疾病相关的序列变异）。有许多类型样品可用做 NGS 检测项目的参考样品，例如来源于人类细胞系的特征性 DNA、合成 DNA 或电子文档。

在目前国家还没有统一标准品制定和发放之前，可以实验室可以自行设计和采用细胞系 DNA、合成 DNA 或电子数据文件等用于检测和分析各阶段的质量控制，实验室应提供自制标准品的参数和方法，以及验证试验的结果。

## 12.检测结果解读与报告

检测报告应满足临床实验室报告的一般要求，包含个人基本信息、检测结果，并在每份报告中注明本次检测采用的方法概要、检测方法局限性、本次检测关键参数（测序覆盖深度等），结果解释所参考的文献资料等信息。

### 12.1 检测结果的临床解读

检测报告应满足临床实验室报告的一般要求，包含个人基本信息、检测结果，

并在每份报告中注明本次检测采用的方法概要、检测方法局限性、本次检测关键参数（NGS 的测序覆盖深度等），结果解释所参考的文献资料等信息。

### 12.1.1 基因的核酸信息

基因的核酸信息包括染色体基因组 DNA 序列和 mRNA 序列，核酸信息参考 NCBI GenBank 核酸序列数据库参考序列（Reference Sequence，RefSeq）。基因组 DNA 序列 GenBank 注册号前面用 NT、NC 或 AC 加下划线进行标注，其中以“NT_”标注的序列为 BAC 克隆或鸟枪测序法获得的不完整的基因组测序序列。如 10 号染色体上的片段 NT_030059，同一序列号有不同的版本号时，后面用点加版本号表示，如 NT_030059.14。成熟 mRNA 转录本序列的注册号前用 NM 加下划线（NM_）进行标注。微小 RNA（microRNA，miRNA）的核酸序列信息参考 miRBase 序列数据库，miRNA 前体序列前用“MI”标注，miRNA 的成熟体序列前用“MIMAT”标注。

### 12.1.2 基因及变异命名

检测报告中的基因名称依据人类基因命名委员会（human gene nomenclature committee，HGNC）1979 年颁布的人类基因命名指南。基因名称和符号参考 HGNC 数据，其主要原则包括：任何一个基因的命名具有唯一性，基因的符号缩写形式可代表对基因名称的概括，基因中只包含拉丁符号和阿拉伯数字，基因符号中不含标点符号，不包含代表基因组的“G”，不包含代表“人类”的字母如“H”或“h”，但人类微小 RNA 基因命名包含代表人类的字母“hsa”。人类基因用大写拉丁字母、斜体表示。

报告中应该列出测序分析发现的所有变异，并根据人类基因组变异学会（Human Genome Variation Society, HGVS）制定并公布的规则进行变异命名。蛋白编码基因中核苷酸的位置一般以编码序列（coding DNA sequence，CDS）翻译起始密码子 ATG 的 A 设为 1，转录起始位点上游一位为-1，翻译终止密码子 3 端第一位命名为*1，后一位为*2，依次类推。对于内含子起始片段内的位点，以上一外显子最后一位核苷酸的位置、加号和内含子内的位置表示，如 c77+1G；对于内含子末端的位置，以下一外显子第一个核苷酸的位置、减号和内含子上游的位置表示，如 c.78-2A。基因突变时，核苷酸替代用“>”（变化为）表示，如

c.76A>C 表示第 76 位由 A 变为 C，c.*46T>A 表示终止密码子下游 3□ 端非翻译区 46 位核苷酸由 T 变为 A。核苷酸缺失是一个或多个核苷酸缺失的序列变异。核苷酸缺失用“del”表示，符号前加上缺失的核苷酸的位置，位置的始末之间用下划线链接，如 g.210_211delTT。复制用 dup 表示，其前面为复制的第一个至最后一个核苷酸的序列。SNP 有参考号的需列出 dbSNP 数据库中的参考号。

### 12.1.3 数据的临床解释

对于测序发现的序列变异需要参照 HGVS 制定的规则进行描述，尽可能根据现有的文献结果给出解释，指明检出的变异是否改变蛋白质编码和功能，与疾病表型之间的关系等信息。对于在测序中发现的变异，建议采用以下的分类进行表述：

（1）序列变异先前曾经有报道，已被公认与疾病相关。序列变异的临床结果之间的关系有比较可行的文献报道支持。

（2）序列变异之前未被报告，但预期会导致疾病。一般来说，如果变异是插入缺失、移码突变、剪接位点 AG/ GT 的突变，都会干扰正常的蛋白质合成和细胞的转录和翻译调节过程。

（3）序列变异之前未被报告，可以与疾病无关，也可能与疾病有关。错义变化、在框（in-frame）插入缺失、剪接位点突变，都可能影响基因表达或细胞内处理过程，可能与疾病相关联。一般需要进一步结合其他检查才能澄清这些变异的临床意义。

（4）序列变异之前未被报告，很可能不会致病。这些变异一般不改变编码序列、碱基变化不会对已知的细胞内信号处理或调节途径发生影响。

（5）序列变异先前曾经有报道，是一个公认的中性变异。有证据表明，该序列变异在正常人群中可以观察到，与疾病发生之间没有相关性。

（6）序列变异不是预期疾病的病因，但有报道与另外一种疾病临床表现有关。一般来说，对于检测中发现的这些与患者预期的临床表型无关的偶发突变（incidental findings），可以不予报道；但是，如果检出的突变可能对患者或其家人带来明确的致病性和危害，则建议在检测结果中如实报告。实验室应参考现有

文献，为特定项目制定偶发突变的报告原则。

总体而言，测序结果的报告应该尽可能参考已有的疾病和/或基因相关的文献报道或疾病诊治指南，经过全面分析后，再作出准确解读和报告。对于文献暂时尚未报道、不确定是否有临床相关性的变异，实验室需要制定相应的标准进行声明，并有后续的处理策略。

**12.2 检测报告**

检测结果以检测报告单的形式发放，需提供纸质版检测报告，有条件的检测实验室可以电子版的形式发放报告，并建立网络查询系统，送检医生通过登陆网站进行检测结果的查询。

检测报告应包含的内容：医疗机构（及科室）名称、项目名称、样品唯一编号、送检医院及科室、受检者姓名、性别、出生日期、样品类型、样品采集时间、样品接收时间、送检医生姓名、疾病诊断、检测结果报告单出具的时间、样品处理过程、检测方法、检测过程、检测结果并附相关图表、结果解释、临床意义和建议、检测的局限、参考文献、检测单位联系信息、检测人员、结果解释人员及报告审核人员签字等。

对于测序结果的临床报告，下列信息对于理解和判读检测结果非常重要，在设计最终的报告模板时，应该尽可能的全部包含：

（1）测序所分析的基因和/或染色体区域，应该采用 HGNC 的基因命名规则明确注明。

（2）所用的参考序列（RefSeq 存取号）应该注明。

（3）分析的基因和/或染色体的具体区域应该注明，例如编码区、外显子、剪切位点等。

（4）对于测序发现的序列变异和确切位点，应该采用 HGVS 的命名规则详细注明。

（5）核酸变异所致的氨基酸编码改变、及其可能导致的蛋白质功能变化，也应该采用 HGVS 规则、并参考文献资料详细列出。

（6）检测结果的临床相关性应该注明。因为有关变异致病性的信息可能随时间而改变，如果检出的变异可能造成潜在的临床效应也应注明。

（7）检测方法的技术局限性应该明确表述。

（8）对于检测结果的解读的局限性也应该包含（例如，检测结果阴性时，不能排除在基因组其他部位还存在致病突变）。

（9）结果分析过程中所用到的所有数据分析软件包、数据库等应该注明名称和版本号。在线数据库最好注明网址。

（10）应该提供一份与患者表现型有关的基因信息摘要，如果建议患者继续检测其他基因，也建议在报告中注明。

（11）如果根据患者的检测结果，推测需要进一步检测其他人员的基因，也应该在报告中做出建议，例如，对于遗传病患者的家庭成员，应该按照遗传病的遗传类型和外显度，对家庭成员提出基因筛查建议。

（12）如果已知某种变异和疾病的治疗或预后之间存在相关性，应该在报告中给出建议。

结果报告需要有严谨有效的发放和审校流程，以确保检验信息的完整、有效、及时、正确、隐私。首先要对检测结果报告进行审核分析，包括审核检测过程的有效性，受检者的基本信息，结果数据分析。审核者应当是主管技师以上的工作人员、本专业实验室负责人、高年资检验人员和临床实验室主任授权人，审核者对检验报告的质量负责。

### 12.3 检测结果回报时间（turnaround time，TAT）

TAT 是指从采集血样样品到报告结果的时间。应该针对每个测序项目的实际检测流程和临床需求，以书面的形式制定合适的结果回报时间。

### 12.4 检测报告的机密性

所有的检测结果均具有机密性。结果可用于指导临床的个体化医疗。如果将结果报告直接告知个人，需要有相应的指导，使其理解检测的结果，了解检测方法的不足。

### 12.5 检测记录的保存和患者报告的可追溯性

实验室需要根据当地卫生行政管理部门要求，明确各种不同测序数据的保存类型和时间。一般检验报告单至少保存两年，检测结果数据至少保存两年；室内质控和参加室间质量评价记录和质控信息至少保存两年；仪器状态和维修记录要保留到仪器使用终身。检测结果的查询通常可根据患者姓名、样品编号、检测项目和送检日期进行查询。检测报告发放后收到检测报告投诉需记录并统计，分析原因，避免二次错误。

测序原始图像数据可以不用长期保存，其他后续数据（例如 BAM、FASTQ、VCF 文件）需要保存一段时间以供后期验证时调用。建议测序原始或早期的数据保存时间不少于 2 年。对于包含序列变异信息的 VCF 文件和包含医学解释的正式报告，应该保存更长时间，建议长期保存。

测序数据可以本地或云计算中心方式保存，但需要采取制定和采用合适有效的措施来保证涉及个人隐私的医学资料的私密性和安全性。

## 13.NGS 检测实验室的评估与准入

NGS 包括两个步骤，（1）湿实验分析过程，包括样品处理、文库构建和测序；（2）生物信息学的数据分析流程，包括序列比对、注释及变异识别。这两个过程在项目建立和整体优化中都是相互关联、密不可分的。NGS 测序平台一次产生的大量测序数据，对于实验室的技术和管理文档建立、结果确认、质量控制和质量保证、数据存储，以及新技术和数据分析软件的评价和采用等过程，都提出了很高的要求。

鉴于 NGS 的技术的先进性及其对于个体化医学和精准医学的转化实践可能带来明确的临床获益，应该为这样的新产品和新技术提供临床准入途径。NGS 技术从上游到下游的测序流程的全部完成，需要确保测序仪器、测序反应通用试剂、测序目的专用试剂以及下游配套的分析软件和算法等多个环节的兼容性和有效性，才能得到准确性、可重复性较高的结果；加之 NGS 检测试剂和耗材非常昂贵，这都增加了 NGS 进行大范围临床试验验证的难度。因此，通过传统的检测试剂盒申报的方式进行临床应用准入，需要比较长的时间周期和昂贵的成本投

入；而且，NGS 技术流程以及数据解读具有非常高的复杂性，即使有个别试剂盒获得批准，在实际应用过程中同样需要强化上市后的质量监管，包括 NGS 实验室的内部质量控制、实验室室间质量评价等全程质量管理工作，才能确保检测结果的准确可靠。因此，对于 NGS 技术在临床应用准入中的监管工作，将主要采用实验室能力评估与准入机制，即通过对符合条件的实验室进行认证和监督，经过认证合格的实验室所可以将自主开发的检测项目（LDT）用于临床检测。

鉴于 NGS 测序技术在检测技术和数据分析等很多方面具有诸多复杂性和不确定性，建议目前只能在 NGS 检测试点单位、采用自主开发的检测项目（LDT）的方式，在通过能力验证的实验室内部开展临床检测服务。

从事 NGS 检测的医学实验室应该接受定期、不定期现场评估，主要评估标准可参照《个体化医学检测实验室管理办法》执行；另外，结合 NGS 的技术和质量管理体系的特点，制定以下具体评估内容，涉及到的内容可以参见附录 C 表 C.1《NGS 检测实验室评估表》。

### 13.1 NGS 检测实验室的资质要求

将 NGS 应用于临床检测的实验室必须是经批准的个体化医学检测实验室和临床基因扩增检验实验室。需要拥有《医疗机构执业许可证》，可以是医院内的医学检验科，或是具有独立法人资格的第三方临床医学检验所。

### 13.2 实验室的设施与设备及整体要求

NGS 检测实验室必须具备相应的专业技术力量，包括实验室空间环境、仪器设备、技术流程、实验操作人员和数据分析人员等。

作为用于临床检测的核酸分析实验室，必须要保证结果的可靠性和准确性，首要措施就是防“污染”。因此，实验室的各工作区的区域设计需要严格按照《医疗机构临床基因扩增检验实验室工作导则》进行。如果 NGS 项目的结果需要经过 Sanger 测序进一步验证，还需要为 Sanger 测序流程额外设置实验区域。

NGS 检测需要的环境条件不同，需要定时监测和记录环境条件。测序实验室的环境温度和湿度对于测序过程和结果的影响非常明显，因此需要特别注意。实验室的各工作区域需保持清洁。需要为工作人员制定相应的标准操作程序，确

保实验样品的安全存放和工作区的清洁，以防止生物污染和扩增产物的交叉污染。

实验室的每台仪器设备均应有相应的使用说明或操作手册、维护、校准或性能检验的程序。实验室的每台仪器设备均应有唯一性标签或其他识别方式。实验室仪器设备的操作人员需经过上岗培训和授权，每次使用后有完整的仪器状态记录。实验人员应经过严格培训，并随时关注有关仪器设备的最新使用说明书。

使用计算机或自动化设备收集、处理、记录、报告、存储或检索数据时，实验室应确保做到：针对每台仪器设备使用的计算机软件，制定详细的使用说明和操作流程，并验证其适用性；制定并执行相应程序，保护计算机程序和代码，以保护资料的完整性，防止无意的或未经授权者访问、修改或破坏数据；应维护计算机和自动化设备，以确保其正常运转，并应提供相应的环境和操作条件。

实验室应制定不同污染废弃物的处理制度，污染物的最终弃置应符合国家（国际）环境或健康安全规则。

鉴于 NGS 的操作、检测技术和数据分析流程的复杂性，操作人员需要具备娴熟的基因组学实验技能，接受 NGS 技术的理论和实验培训，并经过考核合格后才能上岗；数据分析人员具有编程能力、数据分析和解读能力；对于检测报告中测序结果的医学解释，应具有相关法规及技术要求的工作经验，以及医学检验、病理、药理或遗传相关专业技术职称。

### 13.3 实验室的质量控制管理体系评估

NGS 检测实验室开展个体化医学检测实验室项目时，所采用的试剂原则上必须通过 CFDA 批准。如果 NGS 实验室是经过授权批准的 NGS 检测试点单位，所使用的试剂可以是自制试剂（LDT），但自主研发过程需要符合《个体化医学检测 LDT 研制技术规范》要求，并将检测项目的临床意义、试剂的性能测试、检测流程和质量控制等相关信息存档和上报。

NGS 检测实验的整个流程，包括检验前准备、测序检验过程、检验后程序及结果报告，均应符合 GBT 22576-2008《医学实验室-质量和能力的专用要求》的规定，并按照《个体化医学检测质量保证指南》的细则要求执行。需要提供实

验室质量管理体系规范文件和操作规程（SOP）；具有严格的室内质量控制措施，以及完善的实验室管理制度和相关技术规范；按要求定期参加实验室室间质量评价或进行有效的实验室室间比对；其他还包括仪器设备维护、客户满意度维护措施等。随着 NGS 技术的不断发展和进步，NGS 检测实验室还应该有持续的质量保证和改进计划和实施措施。

### 13.4 SOP 编写

实验室必须为每一个 NGS 检测项目建立齐全和规范的书面 SOP 文档，SOP 源于一些标准文件和实验室实验工作经验的积累，应包括试剂准备、样品采集、样品接收与预处理、核酸提取、测序方法和参数、测序仪器操作、生物信息学算法/软件包和流程、结果分析和报告、实验室安全措施等临床检验的任何一个环节。SOP 的编写应注意通俗易懂、注重细节、清晰明了、图文并茂。实验室工作人员应严格遵循标准操作程序中的步骤要求进行操作，当发现 SOP 存在问题时，经过技术研发小组的工作人员讨论、实验验证后及时修改并签字确认。

除了符合质量管理体系的格式和内容要求之外，NGS 检测项目的 SOP 中还应该包括以下内容：

- 对于检测项目分析靶区域的描述（例如，多基因组合、外显子组、全基因组、染色体非整倍性等）。
- 描述检测项目可以适用的样品类型（例如外周血、血浆、FFPE 石蜡标本），并指明本项目建立和验证过程中所采用的样品类型。
- 用于核酸提取和文库构建的试剂和方法。
- 用于靶区域（例如多基因组合、外显子组）捕获的方法和试剂。
- 用于湿实验分析过程的对照与设置方法。
- 所用的测序平台、测序反应的试剂或耗材的版本（例如反应池/flow cells、芯片）。
- 测序设备上用于产生原始数据及输出格式（如 fastq 文件）的软件和版本号。
- 用于监控和评估测序过程的指标和质量控制参数。
- 分析流程的接受和拒绝标准，例如，如果测序覆盖度不能达到特定检测

项目的最低要求，采用 NGS 进行靶向区域测序的结果就可能存在假阳性或假阴性，从而与后续的 Sanger 测序验证结果不符，因此就需要对于这类项目建立可以接受的靶向区域覆盖度标准。

对于所有 NGS 检测项目，需要根据实验内容的靶区域的不同，建立合适的确认方法。一般的确认过程包括：

- 描述检测的区域（如外显子、多基因组合）和检测的方法。
- 对于湿实验的全环节进行确认（如样品处理、核酸提取等）。
- 干实验的分析流程和结果准确度进行确认。

对于采用自主研发试剂（LDT）的实验室，检测项目需要具有严格标准操作规程（SOP）、通过室内质控和室间质评。

### 13.5 实验室的检测报告与服务效率

为提高 NGS 检测实验室的服务效率，需要科学、高效地做好实验室的管理工作，包括检测项目的目录准备、样品采集和运送要求，报告时间、远程报告查询、检测报告和结果的解释等，均应符合评估指南要求。

### 13.6 实验室人员培训

NGS 检测实验室管理和技术人员的专业、学术水平、教育培训情况需要符合要求；人员培训不仅包括外部的培训，也需要重视内部培训，定期组织操作人员进行内部课程学习及实验技术操作培训并考核。培训者要制定适宜培训考核计划，每次培训要有详细记录，分析考核成绩。

NGS 实验室的技术人员需持有 PCR 上岗证，而且应接受 NGS 技术的理论与实验培训，培训后经考核合格后方可上岗工作。

NGS 数据分析人员应该参与 NGS 数据分析培训，并考核合格后方可上岗；或已在 NGS 技术服务行业从事数据分析 3 年以上，但需要有详细的接受内部或外部培训的记录或证明资料。

## 附录 A 目前市场主流测序平台的主要性能与技术参数比较

表 A.1 目前市场主流的测序平台性能与技术参数比较

| 代数 | 第 1 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 2 代 | 第 3 代 | 第 3 代 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 平台名称 | Sanger 3500/3730 | Roche 454 FLX 系统 | Roche 454 GS Junior | HiSeq 2500 | MiSeq | Ion Proton | Ion PGM | PSTAR-IIA | In-house lab-built instrumentation | PacBio RS system | GridION MinION |
| 公司 | Thermo/ABI | Roche/454 | Roche/455 | Illumina | Illumina | Thermo/ABI | Thermo/ABI | 华因康基因 | BGI/CG | PacBio | Oxford Nanopore |
| 测序原理 | Sanger/毛细管电泳测序 | 焦磷酸测序 | 焦磷酸测序 | 可逆链终止物和合成测序 | 可逆链终止物和合成测序 | 合成测序 | 合成测序 | 连接测序 | 复合探针锚杂交和连接技术 | 单分子实时测序（SMRT） | 蛋白纳米孔外切酶测序 |
| 检测方法 | 荧光/光学 | 光学 | 光学 | 荧光/光学 | 荧光/光学 | 检测 pH 值 | 检测 pH 值 | 荧光/光学 | 荧光/光学 | 荧光/光学 | 电流 |
| 准确性 | 99.9% | 99% | 99% | 98% | 98% | 98% | 98% | 99% | - | 87% | - |
| 读段长度 | 600-1000bp | 230-500bp | 500bp | Rapid 2x150bp<br>High 2x125bp | 2 x150-2x300bp | 150-200bp | 35-400bp | 20-50bp | 10bp | 3-10kb | - |
| 单次检测通量 | 100-600Kb | 200Mb | 50Mb | Rapid 10-180Gb<br>High 50Gb-1Tb | 300Mb-15Gb | 10-100Gb | 3Mb-2Gb | 1-2Gb | - | 250Mb | - |
| 单次运行时间 | 20 min-3 h | 24 h | 24 h | Rapid 7-40h<br>High <1d-6d | 5-55h | 2-4h | 2-8h | 8 h | - | <24h | - |
| 优点 | 读段长，准确性高，易识别碱基重复区和多聚序列 | 读段长；比第一代的测序通量大 | 测序通量低，易于临床灵活设计检测项目 | 测序通量高，平均试剂成本低 | 测序通量低，易于临床灵活设计检测项目 | 无光学系统；成本低。1d 内可完成人全基因组测序全流程。 | 无光学系统；成本低。可在 1d 内完成；通量低，适于临床检测 | 测序通量低，易于临床灵活设计检测项目 | 测序通量高，平均试剂成本低；每个测序步骤独立，错误累积降低 | 不需扩增，读段长，易发现结构变异；可直接检测甲基化 | 有望获得长读段，无需荧光标记 |
| 局限性 | 通量低；价格高 | 碱基重复区识别困难；操作复杂；仪器昂贵 | 碱基重复区识别困难；操作复杂；仪器昂贵 | 读段短；仪器昂贵，测序及数据分析时间长 | 读段短；仪器、试剂昂贵，测序及数据分析时间长 | 读段较短；碱基重复区识别困难 | 读段较短；碱基重复区识别困难 | 读段短，数据分析时间长 | 读段短，样品制备复杂，无商业化仪器 | 准确率低，应用成本高，仪器昂贵 | 切断的核苷酸方向可能误判，工艺待改进 |

## 附录 B 面向公众开放的 NGS 数据分析软件包

表 B.1 面向公众开放的 NGS 数据分析软件包

| 软件工具 | 网址 |
|---|---|
| **Short Read Mapping（短读段映射）** | |
| Bowtie | http://bowtie.cbcb.umd.edu |
| BWA | http://bio-bwa.sourceforge.net/ |
| SOAPaligner/soap2 | http://soap.genomics.org.cn/soapaligner.html |
| Maq | http://maq.sourceforge.net/ |
| RMAP | http://rulai.cshl.edu/rmap/ |
| SeqMap | http://www-personal.umich.edu/~jianghui/seqmap/ |
| SHRiMP | http://compbio.cs.toronto.edu/shrimp/ |
| SSAHA2 | http://www.sanger.ac.uk/resources/software/ssaha2/ |
| SNAP | http://snap.cs.berkeley.edu/ |
| **Variation Call（变异识别）** | |
| Cortex | http://cortexassembler.sourceforge.net/ |
| SNVMIx | http://compbio.bccrc.ca/software/snvmix/ |
| ExomeCNV | https://secure.genome.ucla.edu/index.php/ExomeCNV_User_Guide |
| GATK | http://www.broadinstitute.org/gatk/ |
| SAMtools Pileup | http://samtools.sourceforge.net/pileup.shtml |
| STRiP | http://www.broadinstitute.org/gatk/guide/topic?name=third-party-tools |
| BreakDancer | http://gmt.genome.wustl.edu/breakdancer/1.2/index.html |
| TIGRA-SV | http://gmt.genome.wustl.edu/tigra-sv/0.1/index.html |
| **Annotation Analysis（注释分析）** | |
| idb2.R | http://compbio.charite.de/contao/index.php/ibd2.html |
| Genome MuSiC | http://gmt.genome.wustl.edu/2012/04/27/genome_music_0.3_released.html |
| mutationSeq | http://compbio.bccrc.ca/software/mutationseq/ |
| SnpEff | http://snpeff.sourceforge.net/ |
| SomaticSniper | http://gmt.genome.wustl.edu/somatic-sniper/current/ |
| **Splice Junction Detection（剪切位点检测）** | |
| HMMSplicer | http://derisilab.ucsf.edu/index.php?software=105 |
| MapSplice | http://www.netlab.uky.edu/p/bioinfo/MapSplice |
| SOAPsplice | http://soap.genomics.org.cn/soapsplice.html |
| SpliceMap | http://www.stanford.edu/group/wonglab/SpliceMap/ |
| Supersplat | http://mocklerlab.org/tools/1 |
| TopHat | http://tophat.cbcb.umd.edu/ |
| **Isoform Expression Quantification（可变剪切的表达定量）** | |
| ALEXA-Seq | http://www.alexaplatform.org/alexa_seq/index.htm |
| Cufflinks | http://cufflinks.cbcb.umd.edu/ |

| 软件工具 | 网址 |
| --- | --- |
| IsoInfer | http://www.cs.ucr.edu/~jianxing/IsoInfer.html |
| MISO | http://genes.mit.edu/burgelab/miso/ |
| MMSEQ | https://github.com/eturro/mmseq#mmseq-transcript-and-gene-level-expression-analysis-using-multi-mapping-rna-seq-reads |
| rSeq | http://www-personal.umich.edu/~jianghui/rseq/ |
| Scripture | http://www.broadinstitute.org/software/scripture/?q=home |
| **Differential Gene Expression（差异基因表达）** | |
| baySeq | http://www.bioconductor.org/packages/2.8/bioc/html/baySeq.html |
| Cuffdiff | http://cufflinks.cbcb.umd.edu/ |
| DEGseq | http://bioinfo.au.tsinghua.edu.cn/software/degseq/ |
| DESeq | http://www-huber.embl.de/users/anders/DESeq/ |
| edgeR | http://www.bioconductor.org/packages/release/bioc/html/edgeR.html |
| GPSeq | http://www-rcf.usc.edu/~liangche/software.html |
| Myrna | http://bowtie-bio.sourceforge.net/myrna/index.shtml |
| NOISeq | http://bioinfo.cipf.es/noiseq/doku.php?id=start |
| ASC | http://www.stat.brown.edu/Zwu/research.aspx |
| GENE-counter | http://changlab.cgrb.oregonstate.edu/node/26 |
| **Transcript Reconstruction（转录子重建）** | |
| Cufflinks | http://cufflinks.cbcb.umd.edu/ |
| Scripture | http://www.broadinstitute.org/software/scripture/?q=home |
| Velvet | http://www.ebi.ac.uk/~zerbino/velvet/ |
| Trans-ABySS | http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss |
| Trinity | http://trinityrnaseq.sourceforge.net/ |
| Oases | http://www.ebi.ac.uk/~zerbino/oases/ |

注：随着生物信息学的不断发展，还有更多的免费和商业软件包在不断发布。

缩写：BWA, Burrows-Wheeler Aligner; GATK, Genome Analysis Toolkit; MISO, Mixture of Isoforms; MuSiC, Mutational Significance in Cancer; SHRiMP, SHort Read Mapping Package; SOAP, Short Oligonucleotide Analysis Package; SSAHA2, Sequence Search and Alignment by Hashing Algorithm; SNAP, Scalable Nucleotide Alignment Program; STRiP, Genome STRucture In Populations.

## 附录 C NGS 检测实验室评估表

表 C.1 NGS 检测实验室评估表

| 项目 | 评估内容 | 评估结果 |
|---|---|---|
| 实验室资质 | 医疗机构执业许可[a] | 证书编号：_____<br>有效期：_____ |
| | 临床基因扩增检验实验室验收[a] | 证书编号：_____<br>有效期：_____ |
| | 其他认可或认证 | 证书名称：_____<br>证书编号：_____<br>有效期：_____ |
| 实验室负责人 | 姓名 | |
| | 任职证书或证明 | □有 □无 |
| | 工作性质 | □全职 □兼职 □顾问 |
| 实验室工作人员 | 主要技术负责人简历 | □有 □无 |
| | 高级技术人员 | _____名 |
| | 中级技术人员 | _____名 |
| | 初级技术人员 | _____名 |
| | 生物信息分析人员 | _____名 |
| | 临床结果解读与签发人员 | _____名 |
| | 实验室日常检测工作中是否有高级技术人员指导 | □是 □否 |
| | 现场检查实验室时，是否有指定人员陪同检查 | □是 □否 |
| | 技术人员是否具有相应的专业知识或相关职称 | □是 □否 |
| | 获得专业技术证书的技术人员 | 证书：_____ 人数：___ |
| | 实验室人员是否接受 NGS 技术及数据分析培训[a] | □是 □否 |
| 实验室设施与设备 | 区域设计符合临床基因扩增检验实验室要求[a] | □是 □否 |
| | 检测和分析设备操作、环境监测等作业指导书（SOP）是否齐全与规范 | □是 □否 |
| | 检测和分析设备是否有定期维护并有相应记录 | □是 □否 |
| | 是否建立专用的数据处理和存储的计算机平台 | □是 □否 |
| | 污染废弃物的处理是否符合规定 | □是 □否 |
| 室内质量控制活动 | 实验室质量管理人员 | □全职 __名 □兼职 __名 |
| | 实验室是否有完整的室内质量控制文件体系 | □是 □否 |
| | NGS 检测项目的作业指导书（SOP）是否齐全和规范 | □是 □否 |
| | 实验室研发的项目（LDT）和试剂是否经过专家委员会论证[a] | □是 □否 |
| | 检测和分析系统（试剂、参考样品、质控品、仪器、样品溯源、操作方法、平行确证平台和数据分析流程）是否符合需要[a] | □是 □否 |
| | 每个检验项目是否建立了可报告范围和参考范围 | □是 □否 |
| | 每个检验项目是否建立了覆盖深度等 NGS 专用参数 | □是 □否 |
| | 是否拥有自建或公用的核酸测序结果解读数据库 | □自建 □公用 □无 |
| | 是否对质控结果或数据进行统计和汇总分析 | □是 □否 |

表 C.1 （续）

| 项目 | 评估内容 | 评估结果 |
| --- | --- | --- |
| 室间质量评价 | 是否定期参与能力验证等室间质评工作，或者与其他三家以上实验室进行室间比对 | □是 □否 |
| | 对于室间质评结果是否有纠正与改进措施 | □是 □否 |
| | 是否同意将样品送其他实验进行比对 | □是 □否 |
| 检测报告与服务效率 | 实验室提供的检验项目范围能否客户需求[a] | □是 □否 |
| | 实验室检测的样品量是否能满足客户要求 | □是 □否 |
| | 实验室是否有书面材料说明每个检测项目的采样前的准备、样品采集和不合格样品的拒收标准 | □是 □否 |
| | 实验室是否有书面材料说明每个检测项目的样品的转运或运输要求，包括准备、包装、标识、贮存、取样品时间等 | □是 □否 |
| | 实验室是否有书面文件规定对不适当样品的处理要求 | □是 □否 |
| | 实验室在客户有急需时，是否有特殊程序应对 | □是 □否 |
| | 实验室是否提供每项检验服务的出报告时间的书面承话 | □是 □否 |
| | 实验室是否使用标准化的检验申请、以及结果报告的通信协议或系统 | □是 □否 |
| | 实验室对检验报告是否有审核制度[a] | □是 □否 |
| | 实验室是否有改正和改进检验报告的策略 | □是 □否 |
| | 实验室是否向客户提供咨询服务 | □是 □否 |
| 客户满意度调查 | 实验室是否提供其客户名单供联络或咨询 | □是 □否 |
| | 客户类型 | □医院 □诊所 □体检中心 |
| | 实验室已提供服务的年限 | ____年 |
| | 客户满意度 | □优 □良 □中 □差 |
| | 实验室是否有措施保证客户满意度 | □是 □否 |
| | 因为检测质量问题与客户之间产生法律纠纷 | □经常 □偶尔 □无 |
| 如果标记有[a]的检查项为否定结果时，可终止对 NGS 临床实验室的评估。 | | |

## 参考文献


1） CLSI. Nucleic Acid Sequencing Methods in Diagnostic Laboratory Medicine; Approved Guideline—Second Edition. CLSI document MM09-A2. Wayne, PA: Clinical and Laboratory Standards Institute; 2014.

2） CDC, Laboratory Science, Policy and Practice Program Office(LSPPPO), Next Generation Sequencing: Standardization of Clinical Testing(Nex-StoCT) Working Groups：Next-generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) Workgroup Principles and Guidelines 2012.

3） Rehm HL, Bale SJ, Bayrak-Toydemir P, et.al. , Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15(9):733-747.

4） CAP. Molecular Pathology Checklist 04.21.2014

5） Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. World Wide Web URL: http://www.ncbi.nlm.nih.gov/SNP/

6） The Human Gene Mutation Database: 2008 update. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. Genome Med. 2009 Jan 22;1(1):13. World Wide Web URL: http://www.hgmd.org/

7） Online Mendelian Inheritance in Man, OMIM (TM). Johns Hopkins University, Baltimore, MD. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/

8） Catalogue of Somatic Mutation in Cancer, COSMIC. Sanger Institute. World Wide Web URL:http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

9） Human Genome Variation Society, HGVS. Nomenclature for the description of sequence variants. Prepared by Johan den Dunnen. World Wide Web URL: http://www.hgvs.org/mutnomen/

10） den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet. 2001;109:121-124.

11） den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 2000;15:7-12.

12） Richards CS, Bale S, Bellissimo DB, et al.; Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med. 2008;10(4):294-300.

13） Ellard S, Charlton R, Yau M, et al. Practice guidelines for Sanger sequencing analysis and interpretation. http://www.cmgs.org/BPGs/pdfs%20current%20bpgs/Sequencingv2.pdf.

Accessed February 13, 2014.
14） 全国临床检验操作规程（第三版）
15） 白殿一，《标准编写指南——GB/T1.2-2002 和 GB/T1.1-2000 的应用》，中国标准出版社，2002