

癌症基因组测序方案制定的研究进展*

安云鹤¹ 李宝明¹ 李越¹ 尹玲² 曲俊杰² 苏晓星¹ 武会娟^{1**}

(1 北京市科学技术研究院 北京市理化分析测试中心 北京 100094 2 中国农业大学 北京 100083)

摘要 癌症的基因组测序对于癌症的预防、诊断、预后、治疗以及基础生物学研究都有巨大的潜在的应用价值,正是由于其方向广泛,所以基因组测序的方案制定对于实现特定的研究目标,就显得尤为重要。同时,了解了基因组测序方案制定的规则,也有助于评估如今快速增长的发表文献的正确性和重要性。主要论述高通量测序技术在癌症基因组测序中的实际应用,并讨论癌症基因组测序在方案设计和方法学上如何调整,才能更好地实现特定的研究目的。

关键词 高通量测序 癌症基因组测序 方案设计

中图分类号 R318

DNA 测序技术是近年来,生命科学领域发展最迅速的技术领域之一。DNA 测序不仅为遗传信息的揭示和基因表达调控等基础生物学研究提供重要数据,而且在疾病的诊断、治疗和预防等应用研究中也发挥着重要作用。经过几十年的研究,尤其是功能基因组学、表观遗传学等学科的发展,目前基本上形成了一个共识:基因与包括肿瘤在内的多种疾病存在紧密联系,这种关联的本质是基因组序列与疾病以及个体生理特性的联系。随着单核苷酸多态性与人类基因组单体型计划的深入,碱基序列与人类健康的关系将逐渐呈现出来,大量与疾病及特殊体质相关的基因型或变异将会被发现。由于基因组测序的应用范围非常广泛,其方案制定对于实现特定目标才显得至关重要。本文分析和总结了诸多利用高通量测序技术研究癌症的文章,将着重从测序的方案制定方面进行讨论,以期使科研工作者更好的将高通量测序技术应用到科研工作中。

1 高通量测序技术的发展

基因测序技术的进步得益于基因组学的发展。1980 年,噬菌体 ϕ -X174 被完全测序,成为第一个测定的基因组,也由此标志着基因组学这门学科的出现。1990 年代随着几个物种基因组计划的启动,基因组学

取得长足发展。测序完成的基因组大小由原来的几兆,到几十兆,几百兆,上升到现在的几个吉的数据量。同时,就人类基因组而言,人类大概有 30 亿对碱基,而其中只有大概 5% 的碱基具有一定的功能,在这些有功能的碱基中,只有三千万个碱基是编码蛋白质的。可见在庞大的基因组数据中,只有非常少一部分碱基是研究者真正关心的。而真正引起疾病或表型变化的碱基突变率据估算大约为每三千万个碱基中出现一个突变。最近的“人类 1000 基因组项目”数据更新为:每 8.5 千万个碱基中出现一个突变。综上所述,庞杂的基因组数据迫切需要一种高效的测序手段出现,对其进行解读。于是,高通量测序技术应运而生。

在高通量测序技术产生之初,存在三种测序平台,以 Roche 公司的 454 技术、Illumina 公司的 Solexa 技术和 ABI 公司的 SOLiD 技术为标志。随着测序技术的发展,SOLiD 和 454 技术,由于技术和成本等多方面因素相继退出市场。目前,应用最广泛的第二代高通量测序技术即为 Illumina 公司的 Solexa 平台。基于此平台的仪器设备不断升级,相应的建库试剂也不断更新,尤其近几年, Illumina 推出了与肿瘤基因检测相关的 TruSight 和 TruSeq 系列试剂盒,可以说 Solexa 平台的发展非常迅速,已经应用于基因组学,包括测序和表观基因组学以及功能基因组学研究的许多方面,并且正在向临床医学的应用发展。

上述提到的第二代高通量测序技术,主要依赖

收稿日期:2014-09-20 修回日期:2014-10-23

* 北京市博士后科研活动经费资助项目(2013ZZ-34)

**通讯作者,电子信箱:sunnywhj@126.com

PCR 对待模板进行扩增,所以很难避免 PCR 带来的碱基错配、优势片段扩增造成的扩增不平衡。近两年, Pacific Biosciences 公司又开发出了一种新的高通量测序技术,我们称之为第三代高通量测序技术。此技术又叫做单分子实时技术(SMRT),是对单分子 DNA 进行非 PCR 测序并进行实时读取为主要特征的更新的测序技术。SMRT 测序技术在测序速度和读长方面有着巨大的优势和潜力,为人类从基因水平深入理解疾病的发生、发展、诊断和治疗提供了新的手段^[14]。

2 高通量测序技术在癌症研究中的应用

癌症的发病机理一般是源自遗传的基因突变或获得性的体细胞突变。因此,基因组学是研究癌症必不可少的一个方面。2008 年,利用高通量测序技术第一个急性髓性白血病患者的基因组被测定^[5]。截至目前,国际癌症基因组联盟(international cancer genome consortium, ICGC)已经向科研界公布了超过 1 万个癌症基因组的数据。而 20 年前的人类基因组计划,需要国际合作花费 10 年时间,耗资 38 亿美元才能完成一个人类基因组的测序^[6]。随着测序技术的进步和成本的降低,更多的癌症基因组将被测序完成。这些海量数据的获得,不但大大丰富了癌症基础生物学研究的数据和知识,并且大大增加了在其中寻找有关癌症预防、诊断、预后和治疗信息的机会。

尽管基因组测序已经解决了癌症研究中的很多问题,但在此领域,仍然存在大量问题尚未解决。癌症的基因组测序能够在基因组范围内发现与癌症相关的全部基因,但具体到反映的是与癌症的预防、诊断、治疗还是预后相关的基因,则取决于测序之前的方案设计,包括病例收集、测序类型、测序深度以及分析指标等。比如要揭示个体易感肿瘤基因的遗传多态性,那么选择合适的家系或者特定的病例去测序,不但可以减少病例数,而且更能够说明问题。接下来,我们将从癌症基因组研究的特定目标、研究方法和研究类型三个方面对于测序过程中的方案设计进行讨论。

2.1 癌症基因组测序的特定目标(要解决的特定问题)

到目前为止,大多癌症基因组测序相关的研究都是为了实现以下四个特定目标或其中部分目标:发现驱动突变、识别体细胞突变的特征、表征克隆进化特点、推进个性化医疗。第一个方面,确定哪些突变有可能导致癌症表型变化是癌症基因组测序研究最普遍的

目的。发现驱动突变可以提高我们对癌症基础生物学的理解,并由此推动新的治疗方法的发现和发展。以 *zeste2* 基因(*EZH2*)为例,二代测序发现在淋巴瘤中,*EZH2* 基因在临床上具有非常显著的突变率,从而促使 *EZH2* 基因的功能得以解析,并使其成为一个潜在的治疗靶点^[7];第二个方面,发现体细胞突变的特征也有助于获得更多关于癌症基础研究的知识。研究者可以揭示整个突变和 DNA 修复机制的情况和特征。其中, *kataegis*^[8] 和 *chromothripsis*^[9] 这两种突变类型就是在大量的癌症基因组测序中发现的。第三个方面,表征克隆进化的特点是一个非常重要的方面,尤其是对于癌症治疗,同时通过测序可以在核酸水平揭示克隆进化的特点。例如,多数细胞可能具备相同的固有的药物耐受突变类型,然而其中一小部分细胞或产生了新的突变的细胞会产生获得性的药物耐受情况。如有文章报道,大鼠肉瘤中 *v-Ki-Ras2* 基因的突变导致大鼠对 EGFR 靶向治疗的耐受^[10]。第四个方面,即推进个性化医疗,是目前癌症基因组测序比较清晰的应用方向。个体化用药的目的就是为了有目的的对患者选择合适的药物,合适的剂量和治疗时间,从而降低药物毒副作用。成神经管细胞瘤是深入研究个性化用药的模式肿瘤类型,因为它是一个异源性的肿瘤类型,在整个的生存率和分子特征上来说,个体差别都比较大。而且,积极的治疗方案虽然会降低死亡率,但会极大地升高发病率^[11]。因此,找到合适的适用于积极治疗方案的患者对于提高存活的成神经管细胞瘤患者的生活质量具有重要的作用。除了以上提到的四点,还有很多癌症基因组测序可以解决的问题。比如,与儿童癌症相关的新的胚系变异,可以通过对先证者的后代和其父母的研究发现,并且这些特定目标之间是相通的,例如研究克隆进化同样有助于驱动突变的发现。

2.1.1 发现驱动突变 判断体细胞突变是否对癌症的病理过程至关重要,可以看这个突变是否影响到已知的功能性癌症基因,或者是否影响了参与癌症发生的信号通路中的基因成员。目前,对于癌症基因组测序研究的理解大多是基于已有的知识,并非依赖于统计分析,因此,可以不需要大量样本。但现在已知的与癌症相关的基因非常有限,由权威机构桑格研究所公布的癌症基因普查的结果显示,只有 487 个基因是被公认的,而这对于大量数据的癌症基因组研究是明显不足的。

在不同的病人中,相同癌症表型的研究发现,相关

的基因或信号通路存在相同或类似的突变,这表明这些基因存在分子上的共同进化,从而提供了生物进化和选择的有力证据^[12]。生物信息分析的工具,如 MutSig 和 MuSiC 能够找到明显的重复性出现的突变基因^[13-14]。目前,寻找驱动突变最常用的方法是通过识别重复出现的、具有转化作用的突变位点(如错义、无义突变、剪切位点 SNVs 或编码区域 indels)。另外,启动子区域突变位点的负调控作用,可以降低非翻译区或内含子中突变基因的表达,这些在非编码区域明显反复出现的突变基因也表明非编码突变对肿瘤的病理发生也可能是有作用的^[13, 15-16]。同义突变在转录水平、蛋白结构和剪切等方面也具有潜在的调控作用^[17-18]。如果这些突变机制能够失活、激活、缓解或改变基因的功能,研究者应该全面的评估是否结构变异、转录水平或表观遗传学水平的负调控也会影响重复出现的突变基因。

确定了重复性之后,突变模式也可以预示预测驱动突变可能的功能。例如,突变热点(也就是突变反复出现的单个 DNA 位点)和体细胞突变(基因上某一个特定区域成簇的突变),预示着这些基因的功能上调或者是具有促进癌症发生的功能^[16]。相反,分散在一个基因中的重复突变的出现预示其肿瘤抑制的作用^[19]。生物信息的分析同样可以筛选到重复出现的驱动突变基因。

信号通路分析是发现驱动突变的另一个方法,特别是在异源性突变较多,发现很少重复出现突变的癌症类型中,这种方法尤其重要。因为,在癌症的病理过程中重要的突变基因在某些个体中可能是不同的,但不同的突变基因可能具有类似的生物学功能^[13, 20]。另外,有些突变基因在某一个病人中,会与其它基因一起共同作用下调某一个信号通路^[21-22]。在这两种情况下,信号通路分析就可以揭示感兴趣的驱动突变基因,以便于后续研究。但需注意,不是所有突变基因都能网络进信号通路分析,而且,只了解基因产物之间的功能是不全面的。同时,了解相互排斥的突变基因有助于区分不同的癌症亚型^[23],还可以揭示具有类似原癌基因功能的不同的突变位点^[14, 24]。目前,已经有相关的生物信息学软件可用于评估相互排斥的突变基因^[25]。

2.1.2 描述克隆进化 有三种基于二代测序的方案设计,用于阐明癌症的克隆化和分子进化。这些方案设计包括超高深度的重测序^[5, 15, 26],多个区域的测

序^[27-28]和测序多重复样品^[26, 29],它们相互之间不是互相排斥的。第一种,对选择的突变基因进行超高深度的重测序(一般是大于 100X),这使得研究者可以更加精确的评估等位突变基因的突变频率和检测低频突变^[30]。基于等位基因突变频率的成簇分析可以揭示亚克隆的数目、肿瘤细胞内部的异源性^[15, 31],并且可以构建进化树推测亚克隆之间的进化关系^[26, 29-30]。该方案的优势在于它仅需要测序一个样品;第二种,多区域测序也可以揭示肿瘤内部克隆的异质化和实体瘤中细胞的进化情况,无需关注等位基因的突变频率^[32];第三种,测序多个重复样品是基于在一段时间内,在相关的肿瘤细胞中,观察到突变等位基因的突变频率的变化。该方案中除了原发性肿瘤,复发肿瘤^[33-34]和次级转移^[27, 35]也需要被测序。突变频率发生改变或对于继发性肿瘤具有唯一作用的体细胞突变对疾病进展和获得性药物耐受性是非常重要的。多个重复样品进化分析的主要问题之一是测序样品的伦理问题。对于血液癌症来说,这个问题还不突出,因为监测病情进展需要频繁的抽血,从而方便取样用于实验。然而,对于大多数实体瘤来说,获得活检、切除的原发或转移组织却并非易事。

2.2 癌症基因组测序的研究方法(方法学要素)

目前,癌症基因组测序已经有比较成熟的研究方法,其中包括但不限于相对应的正常组织的基因组测序(有时是癌旁组织的基因组测序)。如果要筛查单个核苷酸的变异,运用重测序对获得性变异进行确认实验,至少要保证 30 倍的覆盖度。接下来我们将从匹配的正常基因组、研究 SNVs 的覆盖度、双末端读取检测结构变异、重测序验证等方法学必需的几个要素进行讨论。

2.2.1 匹配的正常基因组 在癌症基因组测序中,排除其相对应的正常组织中的基因变异,有助于发现癌症专有的体细胞突变。血液肿瘤类型的正常组织常常取自皮肤组织^[36-37],而实体瘤的研究经常取外周血单核细胞作为正常基因组的来源^[22, 31]。这种选择可能会引入循环肿瘤 DNA 或循环肿瘤细胞的污染。如果患者必须行手术治疗,手术边缘组织和最接近的淋巴结也可以作为正常基因组的来源^[38],这种方法的创伤最小。但也必须注意,这些组织中除了正常组织,也包括一些残余的疾病细胞,起始肿瘤的早期突变和已经改变的转录组或表观基因组。除了相应的正常基因组来源要严格,生物信息学分析也是允许低水平的污染存

在的。在人类基因组中,平均每个人会遗传到3~4百万个SNP。与之相比,每个成人的基因组中只有几千到几万个SNV是与癌症相关的^[13, 39]。为了验证这些SNVs,研究者必须了解相应的正常组织中大量存在的遗传而来的SNPs。评估正常组织中SNP calling的标准有:这些SNPs能够比对到美国国家生物技术信息中心(US national center for biotechnology information) SNP数据库中的比例,此数据库中的SNP是普遍存在的简单基因突变;这些SNPs从转换到颠换的比率(一般整个基因组的比率为2.1);与相应的SNP基因芯片的一致性。SNP芯片可被用于评估假阴性的比率,但这首先要假设芯片的结果是金标准。

基因组中的突变类型,除了最常见的SNVs这种体细胞突变外,还有结构突变,如拷贝数变化(CNVs)、导致异源性丢失的中枢区域的位置、倒置和置换,也可以通过与癌旁组织的基因组测序比较而定。

2.2.2 研究SNVs的覆盖度 对于有3G数据量的人类基因组而言,得到90G的数据量,覆盖度可以达到30倍,对于研究遗传而来的SNPs是足够的。而正常组织中要达到最优的覆盖度是比较有挑战性的,因为影响SNPs检测的因素很多,如文库的构建方法,测序过程,测序片段的长度,测序片段的比对法则以及生物信息分析的工具。但是如果非常确定的检测遗传表型,一般要测到50倍的覆盖度。

尽管癌症起源于共同的祖细胞,但在癌症的发生发展过程中,经历了克隆增殖,体细胞突变和选择等过程,因此来自同一个病人的癌症细胞,其突变类型也是不一致的。而且,癌症自身的共性,如非整倍体,非癌细胞的污染,和大量不平衡的结构变异都会增加突变等位基因筛查的可变性。这些因素导致,获得性的SNVs与存在于B细胞,通过遗传而来的SNPs不同。遗传性的SNVs相对简单,如果基因是纯和的,那么突变率是0%;如果基因是杂合的,那么突变率是50%;如果基因是纯和突变,那么突变率是100%。获得性的SNVs,其突变率则是连续的。因此,目前30倍的覆盖度对于降低低频突变的假阴性很可能是不够的。事实上,为了检测低至1%~2%的低频突变,更高的基因组测序深度(400~500倍)是必需的。

为了最大程度的降低假阴性,国际癌症基因组协会指出,每个样品的肿瘤细胞要有至少60%~80%是存活的。有些癌症基因组测序研究为了尽可能的降低非肿瘤细胞的污染,会通过微切割^[35],细胞分选^[40],建

立低传代细胞系^[41]和异体移植^[42]等方法处理样品。这些方法对于一些基质含量高的肿瘤类型来说是必需的,如胰腺癌,或者正常细胞的含量较高的肿瘤类型,如血液肿瘤。同时,提高覆盖度可以补偿肿瘤纯度低的情况,并且在有些情况下,这种是最为直接有效的方法。

2.2.3 双末端读取检测结构变异 采用双末端测序癌症基因组,超过30倍的覆盖度就可以检测到详细的基因组特点和SNVs,Indels和结构变异。少于30倍的测序数据也可以检测到结构变异的信息。这些研究可以解决的问题有:在单核苷酸的水平解析结构变异的模式^[34, 43];描述单个癌症患者不同类型的结构变异的分布^[44];检验不同患者之间不同的变异模式^[45];寻找结构变异的进化^[34, 41];发现嵌合突变的基因^[46]。

结构变异可以通过以下三种方法进行预测:(1)可以通过双端测序reads比对到基因组序列时的空间距离和定位进行预测^[47],比如一端reads比对到了人的某一条染色体上,而与之相对应的另一端reads比对到了另外一条染色体上,这就是出现结构变异的可疑位点;(2)一条reads只有部分序列可以比对到参考序列上^[48],这也是可疑位点;(3)利用生物信息学分析软件如Abyss^[49],对测序数据进行从头组装,然后再与参考序列进行比对,结构变异也可以被检测出来。短DNA文库片段(几百个碱基的文库大小)的双末端测序,可以检测到染色体内较小片段的重排^[50]。相反,大片段DNA文库(几千个碱基的文库大小)的双末端测序,可以检测到复杂DNA区域的重排,如重复序列和完全相同的序列区域。并且与小片段文库相比,较少的测序数据就可以达到相同的物理覆盖度^[44]。比较癌症基因组与相应的正常基因组不同的覆盖度是特异性检测CNVs的一种方法,这种方法不依赖于双末端测序。然而,双末端测序提高了reads比对到参考序列的效率。在克隆的异质性程度较高或肿瘤的纯度较低的情况下,CNVs的检测会比较困难,然而,生物信息学的分析可以弥补这一缺陷。

基因组中的结构变异会导致基因的扩增,丢失,打断或重排。然而,转录组测序可用于识别基因组中的结构变异如何改变此基因的转录过程。而且,转录组测序也可以被用于识别转录的嵌合基因。嵌合型蛋白是理想的可用于药物治疗用的大分子靶标,目的是抑制特定的突变蛋白,在降低毒性的同时提高药物的药效。最成功的药物治疗的分子靶标就是用Imatinib(伊

马替尼)抑制了嵌合型的原癌基因 BCR-ABL^[51]。

2.2.4 重测序验证 重测序验证就是用不同的技术验证试验,以最大程度的降低由于特定技术的系统误差所造成的体细胞突变的假阳性,如文库构建的质量不同,测序错误和偏好性,比对的不准确性等。重测序验证的目的就是进一步确认候选突变基因在正常基因组中确实不存在,而在癌症基因组中确实存在。验证实验是为了确定假阳性率,这在研究中是非常重要的。但是每个患者可能会有成千上万的候选突变位点,每个位点都进行确认是不现实的。在癌症患者的基因组中很多突变位点是孤立的,从价格和测序量来看,如果完全用重测序去验证的话,造价太高。以下几种方法可以验证假阳性率同时可以降低验证基因的数量:选择对蛋白结构、功能或表达水平最有可能发生影响的突变,如非同义 SNVs^[52];随机选择体细胞突变的样品^[42];选择感兴趣的突变位点^[33]。国际癌症基因组协会建议,从计算的验证率进行推断,每个样品中捕捉到的突变位点至少有 95% 是真实的,5% 的误差范围需要至少 384 个突变位点被拿出来去进行验证。

一般验证实验即通过 PCR 扩增带有 SNV、Indel、结构变异的候选目的基因,然后通过 Sanger 法进行测序^[5]。其他的验证方法还有质谱^[21, 32]和利用其他高通量测序平台进行目的捕获测序^[31, 36]。但是, Sanger 测序和质谱不能分辨出低频突变^[53]。基因的扩增或丢失可以通过评估测序结果中的 CNVs 与基因组^[35, 43]或 SNP 芯片^[31, 37]杂交结果的一致性来进行验证。需要说明的是,与芯片技术相比较,测序技术在检测更小的 CNVs 时更具优势。

2.3 癌症基因组测序的研究类型

目前,利用二代测序技术测序一个人的癌症基因组价格在 5 000 美元左右,而且测序价格还在下降,因此,使广泛地利用二代测序进行基因组学研究成为可能,但是测序价格仍然是一个不容忽视的问题。方案设计对于最大程度的降低价格、最有效的实现实验目的至关重要。

2.3.1 单样本研究 单样本的研究大多是源于设想,并且对临床实践具有潜在的指导意义,但是这些发现不具有普适性。研究者可以从候选基因中推测哪些基因对于个体癌症患者的病理过程起重要作用。这些推测的理论大部分来源于文献^[5, 36]或系统进化分析^[32]。因此,在创新性和总结的力度上都有所限制。单样品癌症基因组的二代测序有利于阐述此癌症类型的突变

特点^[54]和克隆进化^[31, 52]。这种单样品的基因组测序最适用于个体化基因组用药^[55]以及协助医生对治疗做出决定。

2.3.2 基因组的群体研究 群体研究指相同的癌症类型,或相同的癌症亚型,对这样一群样本进行二代测序。这些研究对于检测基因和信号通路中基因突变发生的几率具有潜在的指导意义。不同癌症患者的复发情况可以很好地证明,有些突变可能参与了癌症的病理过程,但这些不是确定的证据,因为带有致病基因丢失的家系不均衡性也会导致未参与病理过程的临近基因的重复丢失^[56]。同时,二代测序群体癌症基因组还能够发现体细胞突变的特征,或此癌症类型或癌症亚型克隆进化的模式^[15, 30, 45]。

群体癌症基因组测序的目的在于可以无偏的发现一些新的基因,产生一些新的设想。群体基因组测序的统计学意义取决于病例数、肿瘤之间基因组的异质性、目的基因的突变率等等。但是往往由于病例数有限,限制了群体基因组测序的统计学意义。如果研究的目的是找到特定癌症类型或者亚型中突变率较高的基因,国际癌症基因组协会推荐如果检测低至 3% 的突变率,至少需要 100 对正常和肿瘤的配对组织,以及 400 例的验证组织。当然,这种双重设计需要所有发现的体细胞突变基因都在验证试验中被验证。如果突变位点或突变基因以相当高的频率出现,那么大部分的发现基因就不必要再被验证^[57],但是如果发现的突变基因太少,就达不到特定癌症类型整体基因突变情况的高度敏感性。

2.3.3 多组学研究 多组学研究是利用二代测序技术测序一组相同类型或亚型癌症的基因组、外显子组或转录组。特别是,研究者可以测序少量样本的基因组和大量样本的外显子组或转录组,以最大程度的节省成本,并在不同的层面研究感兴趣的问题或基因。当然,每个样本都做几种组学研究也是没有必要的。

外显子测序技术被广泛的用于癌症基因组测序,并且这种方法发现了很多令人兴奋的结果。目前外显子测序技术能够捕获高达 70Mb 的外显子,非编码 RNAs 和具有很强调控潜能的非编码区域。定制外显子测序也可以捕获特定区域。理论上来说,外显子测序不会促进整个基因组中特定突变类型的发现,它只能检测编码 SNVs, Indels 和结构变异^[58]。事实上,单个样品测序价格的降低使得研究者可以关注基因组中的一个亚类,这样可以测得更多的数据获得更高的覆盖

度,从而增加捕捉低频突变的敏感性^[59]。然而,在覆盖度较低或目标捕获探针需要进一步提高的情况下,外显子测序会遗漏一些编码突变^[54]。

转录组测序可以发现 SNVs^[19], indels, 嵌合型转录组^[60], 新的转录本, 可变剪切^[15], 等位基因不平衡^[23]和差异表达的转录本^[61]。应该注意的是, 对转录水平起负调控从而降低转录水平或转录本稳定性的 SNVs 在转录组测序中是捕捉不到的。并且, microRNA 在多数 RNA-seq 中也是捕捉不到的, 需要单独建库测序。与芯片技术相比, RNA-seq 有很多优势, 因为后者是一种数字技术。例如, RNA-seq 在比较不同基因、样品、实验、时间点和处理条件的转录组水平都有提高。而且, 其能够反映一个更加动态的范围, 以及由测序深度所决定的敏感性。但是, 由于受到相应正常组织收集的困难限制, 转录组测序也是有局限的。例如, 如果起源的细胞类型不能得到或者未知, 比较癌症组织与起源细胞的差异表达分析就是不可能的。

多组学的研究优势是两方面的。首先, 由于外显子和转录组的覆盖度所需要的测序量比基因组测序要少很多, 多组学的研究最大程度降低了测序成本, 时间和病例资源。外显子和转录组测序的低价使得研究者可以测序大量样本, 这有利于发现大量重要的重复出现的突变基因。高影响力指的是很可能会影响蛋白功能或表达的体细胞突变。同时, 少数研究发现, 基因组测序能够发现在外显子组和转录组测序中发现不了的、高频重复出现的结构变异、SNVs 和 Indels。其次, 多组学研究的另一个优势是, 一体化的分析, 这有助于寻找用不同的组学方法得到的不同类型的突变如何集中到同一个突变位点, 基因或通路。这是非常重要的, 因为没有任何一种组学方法可以将癌症病理过程中的基因全部捕捉到。癌症基因组阿特拉斯研究网络, 曾经从各个方面对癌症起源的分子原因进行了整合性分析^[46]。生物信息学分析软件, 如 PathScan^[25] 和 PARADIGM^[62] 都可以对多组学的数据进行整合分析。

其中, 寻找多组学之间的相互关系也是癌症基因组测序研究新兴的一个领域。从技术上来讲, 这种方法就是针对同一个样本, 分析不同组学技术之间的整体关系。需要考虑的是, 一种组学技术检测到的基因突变或异常调控如何影响另一种组学数据。事实上, 探寻组学数据之间的相互关系受到以下几个因素的限制。首先是目前, 只有少数研究中, 每个样品用到了不只一种二代测序的组学技术; 其次是缺乏相应的分析

软件, 部分原因是由于研究者在研究不同组学数据之间相互关系时, 提出的问题太多, 一时无法满足。比如, 在研究基因组和转录组相互关系中, 研究者已经了解了转录组数据中 DNA SNVs 的比例, 发现了重新编码氨基酸序列的癌症特异性 RNA 剪切事件, 发现了影响转录本长度和结构的剪切位点突变, 发现了在转录水平上的 DNA 体细胞突变。但很少癌症基因组测序研究揭示了多种表观基因组学和基因组水平基因突变或转录组异常调控之间的关系。此外, ChIP-seq 是在基因组范围内研究组蛋白修饰和 DNA 之间相互关系的非常好的方法^[63]。甲基化的二代测序技术和相应的分析软件也已经出现^[63]。

2.3.4 验证和扩展研究 二维的研究设计包括通过一些样本的研究产生某些设想(这些样本包括单样本研究、基因组学研究和多组学研究), 然后用验证试验在更大的样本群中对这些设想进行验证。验证试验的方法包括: 验证特定核苷酸改变的基因型; 测序单个外显子, 编码区域或者整个基因; 特定捕获成千上万的兴趣基因, 然后用超高深度的测序进行验证。当验证样品的结构变异时, 扩展试验一定是用公众认可的全基因组的芯片数据。

验证试验的目的是验证新发现的这些基因的重现性^[24, 53]。一些癌症基因组测序研究还有另外的目的, 如调查普及性^[57], 评价感兴趣体细胞突变基因和信号通路的临床重要性^[20]。这些已经超出了验证试验的目的, 是对研究的扩展。如果验证或扩展试验由不同的癌症类型组成, 研究者需要明确特定癌症类型普适性的范围。特别是, 为了达到主要目的, 研究者可以在相似的癌症类型(如非霍奇金淋巴瘤和急性淋巴细胞白血病^[19, 64])、有相同特征的癌症类型(如儿科肿瘤^[57])、具有相反特点的癌症(如脑干和非脑干肿瘤^[57])或在各种公共类型的癌症^[34]中, 研究癌症基因组的普适性。多数癌症类型共有的体细胞突变基因或信号通路很可能位于癌症病理过程中的特征表型。然而, 特定癌症亚型共有的体细胞突变对于不同的表型特征、优化的治疗或发现生物标记物也是非常重要的。

3 未来发展

个体化用药将是癌症基因组测序方案设计获益最多的领域。目前, 癌症基因组的二代测序技术用于用药指导还处于初期阶段, 但是在单个病人的用药方面是非常有前景的。尽管如此, 从针对单一病人到针对

群体患者的个性化用药研究仍然具有巨大的潜在价值。最近的研究发现,20%的三阴性乳腺癌^[52]和60%以上的肺癌^[61]患者都有可用于临床的潜在药物治疗靶点。但是,基于群体的个性化用药研究还将面临许多挑战。如果一个随机试验验证了癌症基因组测序导向的治疗方案的安全性和高效性,那么在这个治疗方案中需要用到很多不同的药物,但是具体要用到哪些药物可能会受到实验资助者的影响。如果一个随机试验验证了一种新的治疗方案的安全性和高效性,并且为了符合实验标准,所选择的患者都需要有特定的突变基因、突变的信号通路或突变的特征,那么大量潜在的受试者将需要被严格筛选。此外,所涉及的伦理问题,使得个性化用药的方案只能用于对标准治疗方案失败的终末期患者。因此,这种治疗方案最初只能在最需要和最具挑战性的患者中进行测试。尽管面临众多挑战,二代癌症基因组测序技术正在迅速走向临床应用。

参考文献

- [1] Chin C S, Sorenson J, Harris J B, et al. The origin of the Haitian cholera outbreak strain. *The New England Journal of Medicine*, 2011, 364: 33-42.
- [2] Loomis E W, Eid J S, Peluso P, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Research*, 2013, 23: 121-128.
- [3] Rasko D A, Webster D R, Sahl J W, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *The New England Journal of Medicine*, 2011, 365: 709-717.
- [4] Smith C C, Wang Q, Chin C S, et al. Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature*, 2012, 485: 260-263.
- [5] Ley T J, Mardis E R, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 2008, 456: 66-72.
- [6] International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. *Nature*, 2004, 431: 931-945.
- [7] McCabe M T, Ott H M, Ganji G, et al. EZH2 inhibition as a therapeutic strategy for lymphoma with EZH2-activating mutations. *Nature*, 2012, 492: 108-112.
- [8] Nik-Zainal S, Alexandrov L B, Wedge D C, et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 2012, 149: 979-993.
- [9] Stephens P J, Greenman C D, Fu B, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 2011, 144: 27-40.
- [10] Misale S, Yaeger R, Hobor S, et al. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*, 2012, 486: 532-536.
- [11] Northcott P A, Jones D T, Kool M, et al. Medulloblastomics: the end of the beginning. *Nature Reviews Cancer*, 2012, 12: 818-834.
- [12] Castoe T A, de Koning A P, Pollock D D. Adaptive molecular convergence: molecular evolution versus molecular phylogenetics. *Communicative & Integrative Biology*, 2010, 3: 67-69.
- [13] Fujimoto A, Totoki Y, Abe T, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nature Genetics*, 2012, 44: 760-764.
- [14] Berger M F, Hodis E, Heffernan T P, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*, 2012, 485: 502-506.
- [15] Shah S P, Roth A, Goya R, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 2012, 486: 395-399.
- [16] Chapman M A, Lawrence M S, Keats J J, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*, 2011, 471: 467-472.
- [17] Kimchi-Sarfaty C, Oh J M, Kim I W, et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science*, 2007, 315: 525-528.
- [18] Pagani F, Raponi M, Baralle F E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102: 6368-6372.
- [19] Morin R D, Mendez-Lago M, Mungall A J, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*, 2011, 476: 298-303.
- [20] Molenaar J J, Koster J, Zwijnenburg D A, et al. Sequencing of neuroblastoma identifies chromothripsis and defects in neurogenesis genes. *Nature*, 2012, 483: 589-593.
- [21] Lee W, Jiang Z, Liu J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature*, 2010, 465: 473-477.
- [22] Jones S J, Laskin J, Li Y Y, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biology*, 2010, 11: R82.
- [23] Jones D T, Jager N, Kool M, et al. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 2012, 488: 100-105.
- [24] Cheung N K, Zhang J, Lu C, et al. Association of age at

- diagnosis and genetic mutations in patients with neuroblastoma. *The Journal of the American Medical Association*, 2012, 307: 1062-1071.
- [25] Dees N D, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Research*, 2012, 22: 1589-1598.
- [26] Ding L, Ley T J, Larson D E, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 2012, 481: 506-510.
- [27] Wu C, Wyatt A W, Lapuk A V, et al. Integrated genome and transcriptome sequencing identifies a novel form of hybrid and aggressive prostate cancer. *The Journal of Pathology*, 2012, 227: 53-61.
- [28] Gerlinger M, Rowan A J, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, 2012, 366: 883-892.
- [29] Walter M J, Shen D, Ding L, et al. Clonal architecture of secondary acute myeloid leukemia. *The New England Journal of Medicine*, 2012, 366: 1090-1098.
- [30] Nik-Zainal S, Van Loo P, Wedge D C, et al. The life history of 21 breast cancers. *Cell*, 2012, 149: 994-1007.
- [31] Ding L, Ellis M J, Li S, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 2010, 464: 999-1005.
- [32] Tao Y, Ruan J, Yeh S H, et al. Rapid growth of a hepatocellular carcinoma and the driving mutations revealed by cell-population genetic analysis of whole-genome data. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108: 12042-12047.
- [33] Weiss G J, Liang W S, Izatt T, et al. Paired tumor and normal whole genome sequencing of metastatic olfactory neuroblastoma. *PloS One*, 2012, 7: e37029.
- [34] Ng C K, Cooke S L, Howe K, et al. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *The Journal of Pathology*, 2012, 226: 703-712.
- [35] Turajlic S, Furney S J, Lambros M B, et al. Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Research*, 2012, 22: 196-207.
- [36] Link D C, Schuettpelz L G, Shen D, et al. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *The Journal of the American Medical Association*, 2011, 305: 1568-1576.
- [37] Mardis E R, Ding L, Dooling D J, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *The New England Journal of Medicine*, 2009, 361: 1058-1066.
- [38] Pleasance E D, Stephens P J, O'Meara S, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*, 2010, 463: 184-190.
- [39] Ellis M J, Ding L, Shen D, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*, 2012, 486: 353-360.
- [40] Puente X S, Pinyol M, Quesada V, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 2011, 475: 101-105.
- [41] Campbell P J, Yachida S, Mudie L J, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 2010, 467: 1109-1113.
- [42] Pena-Llopis S, Vega S, Liao A, et al. BAP1 loss defines a new class of renal cell carcinoma. *Nature Genetics*, 2012, 44: 751-759.
- [43] Stephens P J, McBride D J, Lin M L, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 2009, 462: 1005-1010.
- [44] Kloosterman W P, Hoogstraal M, Paling O, et al. Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. *Genome Biology*, 2011, 12: R103.
- [45] McBride D J, Etemadmoghadam D, Cooke S L, et al. Tandem duplication of chromosomal segments is common in ovarian and breast cancer genomes. *The Journal of Pathology*, 2012, 227: 446-455.
- [46] Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 2012, 487: 330-337.
- [47] Korbel J O, Urban A E, Affourtit J P, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 2007, 318: 420-426.
- [48] Onishi-Seebacher M, Korbel J O. Challenges in studying genomic structural variant formation mechanisms; the short-read dilemma and beyond. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 2011, 33: 840-850.
- [49] Simpson J T, Wong K, Jackman S D, et al. ABySS: a parallel assembler for short read sequence data. *Genome Research*, 2009, 19: 1117-1123.
- [50] Fullwood M J, Wei C L, Liu E T, et al. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Research*, 2009, 19: 521-532.
- [51] Druker B J, Guilhot F, O'Brien S G, et al. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *The New England Journal of Medicine*, 2006, 355: 2408-2417.
- [52] Shah S P, Morin R D, Khattri J, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 2009, 461: 809-813.

- [53] Bass A J, Lawrence M S, Brace L E, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature Genetics*, 2011, 43: 964-968.
- [54] Totoki Y, Tatsuno K, Yamamoto S, et al. High-resolution characterization of a hepatocellular carcinoma genome. *Nature Genetics*, 2011, 43: 464-469.
- [55] Demeure M J, Craig D W, Sinari S, et al. Cancer of the ampulla of Vater: analysis of the whole genome sequence exposes a potential therapeutic vulnerability. *Genome Medicine*, 2012, 4: 56.
- [56] Muller F L, Colla S, Aquilanti E, et al. Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature*, 2012, 488: 337-342.
- [57] Wu G, Broniscer A, McEachron T A, et al. Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nature Genetics*, 2012, 44: 251-253.
- [58] Karakoc E, Alkan C, O'Riordan B J, et al. Detection of structural variants and indels within exome data. *Nature Methods*, 2012, 9: 176-178.
- [59] Banerji S, Cibulskis K, Rangel-Escareno C, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 2012, 486: 405-409.
- [60] Roberts K G, Morin R D, Zhang J, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell*, 2012, 22: 153-166.
- [61] Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 2012, 489: 519-525.
- [62] Vaske C J, Benz S C, Sanborn J Z, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 2010, 26: i237-245.
- [63] Hawkins R D, Hon G C, Ren B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 2010, 11: 476-486.
- [64] Zhang J, Ding L, Holmfeldt L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*, 2012, 481: 157-163.

Progress in Cancer Genome-Sequencing Study Design

AN Yun-he¹ LI Bao-ming¹ LI Yue¹ YIN Ling² QU Jun-jie² SU Xiao-xing¹ WU Hui-juan¹

(1 Beijing Center for Physical and Chemical Analysis, Beijing, Beijing Academy of Science and Technology, Beijing 100094, China)

2, China Agricultural University, Beijing 100083, China)

Abstract Discoveries from cancer genome sequencing have great potential application value for cancer prevention, diagnostics, prognostics, treatment and basic biology. Given the diversity of downstream applications, cancer genome-sequencing studies need to be designed to best fulfill specific aims. At the same time, knowledge of second-generation cancer genome-sequencing study design also facilitates assessment of the validity and importance of the rapidly growing number of published studies. Here, we focus on the practical application of second-generation sequencing technology to cancer genomics, and discuss how the study design and method could be adjusted to better achieve the purpose of specific aims.

Key words Second-generation sequencing Cancer genome-sequencing Study design