

·综述·

第二代测序在检测人类基因突变中的应用

周 帆, 林标扬*

(浙江大学 浙江加州国际纳米技术研究院, 中国浙江 杭州 310029)

摘 要: 很多的人类疾病与基因突变有关, 基因突变在疾病的诊断和治疗中起到了至关重要的作用. 第二代高通量测序, 其特点为通量高、速度快、成本低, 给检测基因突变带来了革命性的变化. 该技术检测基因突变的流程简单, 研究人员运用全基因组测序, 目标基因组测序以及转录组测序能够实现基因突变的全方位、高准确的检测.

关键词: 第二代高通量测序; 基因突变检测; 全基因组重测序; 目标基因组测序; 转录组测序

中图分类号: Q755

文献标识码: A

文章编号: 1007-7847(2012)05-0451-06

Next Generation Sequencing Technology and Its Application in Detecting Gene Mutations

ZHOU Fan, LIN Biao-yang*

(Zhejiang California International Nano Systems Institute, Zhejiang University, Hangzhou 310029, Zhejiang, China)

Abstract: Gene mutations can lead to human diseases, and play an important role in diagnosis and treatment of them. Next generation sequencing technology, with features of high throughput, rapid sequencing and low cost, brings revolutionary changes in the fields of detecting gene mutations. The process of using this method to detect mutations is simple. Researchers can combine whole genome resequencing, target sequencing and transcriptome sequencing to detect mutations in all-around, high accurate way.

Key words: next generation sequencing technology; detecting of gene mutation; whole genome resequencing; target sequencing; RNA sequencing

(*Life Science Research*, 2012, 16(5): 451~456)

人类基因组由于其不稳定性, 经常发生基因突变. 基因突变是指 DNA 碱基序列的变化, 包括单核苷酸突变 (single nucleotide polymorphisms, SNP) 和结构变异^[1]. 结构变异是指非单碱基的突变, 包括 DNA 片段的拷贝数目变化及插入、缺失、重复等改变基因剂量的变异, 还包括不改变基因剂量的倒位、基因融合(易位)等. 基因突变是人类多数遗传病、肿瘤等疾病的直接原因, 如大约 50% 的人类癌症中都存在着 *p53* 突变^[2], *APC* 基因与大肠癌的发生密切相关^[3]. 因此, 基因突变的检测具有重要的意义, 传统检测基因突变的方法为核酸分子杂交, 但是此方法过于繁杂, 且容易

对人产生伤害. Sanger 测序法可以发现外显子区域的变异以及拷贝数的变异, 应用此法, 人们发现了很多重要的突变^[4], 但是 Sanger 法通量太低, 不能满足大规模检测基因突变的要求. 上世纪末发展起来的基因芯片技术可以用于大规模检测基因突变, 如 David G. Wang 等用基因芯片一次性检测到了 3 241 个单核苷酸突变^[5]. 基因芯片在应用中普遍存在的问题是其灵敏度低, 利用这种方法不能检测到频率非常低的变异. 2000 年, 人类基因组第一份草图的完成, 开创了测序时代的新纪元, 进入了第二代高通量测序 (Next generation sequencing, NGS) 时代. NGS 方法以高通量、速度

收稿日期: 2012-08-20; 修回日期: 2012-09-18

作者简介: 周帆(1986-), 女, 湖北恩施人, 硕士研究生, 主要从事系统生物学研究; * 通讯作者: 林标扬(1964-), 男, 浙江加州国际纳米技术研究院教授, Tel: 0571-86971772, E-mail: biaoyn@gmail.com.

快、低成本为主要特征,可直接通过聚合酶或连接酶进行体外测序合成. 到目前为止,主要有:以单分子阵列原位扩增测序技术为核心的 Illumina (美国)公司的 Solexa 测序仪器; ABI(美国)公司的 SOLID 测序仪器,它的核心为寡核苷酸链接和检测技术;以及以 emPCR(emulsion polymerase chain reaction)^[6]和焦磷酸测序技术(Pyrosequencing)^[7]为核心的罗氏 454 测序仪. 运用 NGS 方法,能检测到基因组上稀有的突变,而且能够全面地勾勒出基因组上变异图谱,由于其在检测基因突变中的高通量、速度快、灵敏度高等优点,近年来已广泛地应用到检测基因突变领域,本文就其在此领域中的应用进行综述.

1 NGS 检测基因突变的流程

将 NGS 的数据转化为最终的突变包括很多的步骤(见图 1),每个步骤都关系到检测到突变的准确性.

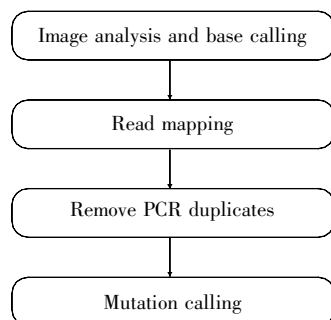


图 1 将 NGS 测序结果转换成基因突变的步骤

Fig.1 Steps for converting raw next-generation sequencing data into a final set of mutations

碱基读取的最大的瓶颈来自于特定长度的序列之间的信号强度差别很大,这主要是由于同一个簇里面的 DNA 的合成并不是同步的所引起,导致了短的插入、缺失的错误率较高,如 Illumina 平台的短插入、缺失的错误率为 1%左右^[8]. 而且

碱基读取的错误率随着读长的增加而增加. 碱基读取错误会导致后续比对、突变读取的错误,因此,减少碱基读取的错误率就尤为重要. 目前,很多的研究机构开发了更加精确的碱基读取软件,如针对 454 平台的 Pyrobayes^[9]; 针对 SOLID 平台的 Rsolid^[10]; 以及针对 Illumina 平台的 Ibis^[11]等. 这些算法比制造商们自行开发的算法减少了 5%~30%的碱基读取错误率^[8]. 虽然如此,但是目前广泛应用的还是制造商们开发的算法.

序列的比对就是将测序得到的短的序列与参考基因组进行比较,找到其在基因组上的位置的过程. 目前,比对工具主要基于两种算法,一是基于“hashing (哈希)”,代表的软件有 MAQ^[12]、Novoalign 和 Stampy^[13],这类工具的特点是比较敏感,目前为止,Novoalign 和 Stampy 总的来说准确率较高;另一种是基于“BWT (burrows-wheeler transform)”,代表的软件有, Bowtie^[14]、SOAP2^[15]和 BWA^[16],这类软件速度非常快,所耗用的内存相对较小,特别适用于比对重复区域.

拥有相同坐标的序列我们就认为它们是聚合酶链式反应(polymerase chain reaction, PCR)冗余,因为这些序列只有非常渺小的可能是随机打碎时产生,这种情况可以忽略不计. PCR 冗余在覆盖度比较低的时候会对突变位点的决定产生非常有害的作用. 它们提高了这些位点的测序覆盖度,在后续突变位点挑选时就会将它们挑出,然而它们其实就是一条序列的贡献非常大而已. 所以,这些序列在进行突变位点找寻之前是应该去掉的^[17].

突变读取软件根据用户设定的碱基质量、测序深度等条件,用自己特定的算法,对突变进行检测,检测到的突变类型不尽相同(见表 1).

2 全基因组重测序

全基因组重测序是指在有该物种参考基因组

表 1 各种突变读取软件能检测到的突变的类型

Table 1 Mutation types detected by the mutation read software

Name	Detected mutation types	Websites
Break Dancer	Insertion, deletion, inversion, gene fusion	http://breakdancer.sourceforge.net/
Pindel	Deletion (1~10 kb), insertion (1~20 bp)	https://trac.nbic.nl/pindel/
MoDIL	Insertion (20~50 bp), deletion (20~50 bp)	http://compbio.cs.toronto.edu/modil/
Variation Hunter	Insertion, deletion, inversion, duplication	http://compbio.cs.sfu.ca/strvar.htm
SegSeq	Copy number variation	http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=182
Samtools	SNP, small indel	http://samtools.sourceforge.net
VarScan	SNP, small indel, copy number variation	http://varscan.sourceforge.net/somatic-calling.html
PEMer	Insertion, deletion, inversion	http://sv.gersteinlab.org/pemer/

的情况下,对该物种的个体进行基因组测序的方法.基于全基因组重测序方法,研究人员能够快速全面地检测出个体基因组上所有的基因突变,从而找到个体间基因的差异.但是,由于人类基因组过于庞大,一次单端全基因组测序很难达到相应的测序深度.如 Illumina HiSeq 2500 平台常规单端测序方法,每次最多能产生 6×10^8 条序列,读长最长为 150 bp,其数据量相当于 30 倍人类基因组.然而,在实际中一次单端测序很可能达不到这样的数据量.目前,公认的挑选 SNP 位点的测序深度应该大于 30^{18-20} .所以,用全基因组测序的方法,如果要得到高质量的突变位点,一般应重复测几次或者用双端测序法.目前,人类基因组重测序已广泛地应用于寻找人类个体间的基因差异,特别是对癌症基因组的重测序,人们发现了很多与癌症相关的变异位点,这对癌症的诊断与治疗提供了更多更加准确的潜在靶点.

Sung Min Ahn 等运用 Illumina 全基因组双端测序,对一个韩国的男人进行了 3 次全基因组双端测序,其中的插入序列长度分别为 100、200、300 bp,总测序深度为 28.95 倍.他们总共发现了 3 439 107 个 SNPs,其中 420 063 个为新的 SNP 位点,通过与中国参考基因组相比较,他们发现了其中 39.87% 的 SNPs 是韩国人特有的.同时他们还发现了 342 965 个缺失在 29 bp 之内,插入在 14 bp 之内的结构变异,以及 2 920 个 0.1~100 kb 的缺失,其中在小于 4 bp 的插入、缺失中,99.5% 跟中国人种是一样的,在 2 920 个 0.1~100 kb 的缺失中 11.3% 为韩国人特有.他们的发现说明,即使如此亲近的种族之间基因组也存在着巨大的差异,因此建立种族特异的参考基因组具有重要意义^[21].

Axel M. Hillmer 等学者用全基因组双端测序(插入序列 10 kb)测了 17 个基因组(5 个乳腺癌原癌组织样本基因组、3 个乳腺癌细胞系(MCF-7、T47D、SKBR3) 基因组、4 个胃癌原癌组织样本基因组、1 个胃癌细胞系 TMK1 基因组、1 个结肠癌细胞系 HCT116 基因组、1 个慢性粒细胞白血病细胞系 K562 基因组、2 个正常人组织基因组)序列,以用来分析上皮癌基因组中的结构变异的特点.通过对这些序列进行结构变异检测,他们发现大部分的插入、缺失、倒位都是在细胞系中存在,而在癌症基因组的组织样本中更多的表达的是重复、染色体间的易位、重组.用此种全基因组

大插入序列的双端测序法能够准确地描绘出基因组上的结构变异,通过对这些变异的研究他们发现,大片段的重复是导致基因组不稳定从而导致上皮癌进一步增殖的原因之一^[22].

3 目标测序

目标测序是指研究人员对感兴趣的特定基因组区域进行测序的方法.研究人员首先对感兴趣的基因组区域进行富集(常见的目标区域富集的方法有 PCR、分子倒位探针(molecular inversion probes, MIP)、芯片杂交捕获(on-array hybrid capture)和液相捕获(in-solution hybrid capture)等^[23]),然后将这些富集的目标区域进行测序.目标测序法由于其测序的目标区域只占全基因组的一小部分,能够达到很高的测序深度,近几年来,已被研究人员广泛应用于外显子测序与特定基因集测序.它们都在寻找特定性状相关的基因变异中起到了至关重要的作用.除此之外,一些基因组上重要位点的测序也在逐渐地增多,比如基因组上 CpG 位点的测序等.研究人员运用此方法,能够更加高效、经济地得到目标区域序列,从而更加准确地得到目标序列的基因变异.

3.1 外显子测序

外显子测序是指将基因组外显子区域 DNA 利用序列捕获技术进行捕获,然后富集进行高通量测序的基因组测序方法.外显子区域仅占人类基因组区域的 1% 左右,但是却拥有约 85% 与疾病相关的变异位点^[24].因此,对外显子区域选择性的测序,然后对外显子区域的变异位点进行检测,并对这些变异位点进行分析,能够更好地理解人类疾病.外显子测序因其测序的高效性、经济性以及能够检测到绝大多数与疾病相关的编码区变异的特点,近年来已广泛地应用于检测编码区基因的突变.

Andreas Keller 等捕获到两个胶质瘤组织及其相应的白细胞中的与胶质瘤相关的 132 个基因的外显子区域,然后用 Illumina Genome Analyzer II 进行双端测序,插入序列的长度为 100~500 bp,测序深度为 308 倍,他们发现了 1 300 多个位于相关基因上的 SNPs,对这些 SNPs 做基因组相关性分析发现,这些突变位点中,有的跟吸烟有关,有的与乳腺癌相关,有的则与更高级别的胶质瘤相关,他们特别注意到在基因 SART1 中的 rs660118 这个与乳腺癌相关的突变位点,在胶质

瘤组织中出现的频率也很高. 另外他们还发现在 21 个与胶质瘤抗原性相关的基因中, 有 20 个有 SNPs, 这为基因变异与抗原的免疫原性密切相关提供了有力的证据^[25].

新生儿糖尿病患者分子诊断对于此疾病的治疗非常重要, 患者如果在基因 *KCNJ11*、*ABCC8* 中有突变, 就可以用口服的磺脲而不是用胰岛素来治疗^[26-28]. 目前新生儿糖尿病病的分子诊断方法是用 sanger 测序法测 *KCNJ11*、*ABCC8* 和 *INS* 基因中的突变, 但是此方法过于繁琐, 而且检测不到频率较低的突变. Amelie Bonnefond 等用 Agilent Sure Select Human All Exon kit 捕获到一个患有新生儿糖尿病的患者所有外显子区域, 该患者在之前的 sanger 法检测中被认为 *KCNJ11*、*ABCC8* 和 *INS* 以及染色体 6q24 区域是没有发生突变的, 然后用 Illumina Genome Analyzer IIx 进行测序检测其突变位点, 发现此患者有一个在 *ABCC8* 基因中的新的非同义突变, 此突变在 348 个参照组中并没有出现, 同样也没有在该患儿的正常的父母、兄弟中出现. 他们认为, 用基因组外显子测序在检测新生儿糖尿病患者所携带的变异时更加全面、快速、经济. 他们建议用此方法作新生儿糖尿病病的分子诊断^[29].

3.2 特定基因集测序

特定基因集测序是指将感兴趣的基因集区域 DNA 利用序列捕获技术进行捕获, 然后富集捕获到的序列进行高通量测序的方法. 很多的人类疾病都与特定的基因相关, 所以对这些基因集进行测序, 找出基因的变异, 分析其致病的机理, 在理解人类的疾病进程中尤为重要. 与外显子测序相比, 特定基因集测序, 加入了特定基因的非编码区域的测序, 非编码区由于其对基因表达的调控作用, 越来越受到科学家的重视. 特定基因集测序, 由于其高效性以及寻找特定基因集变异的全面性, 近年来也受到了科学家们的青睐.

Luca Pagani 等考察了 49 个居住在海拔 2 000 m 的北高加索高地的来自 Avars, Kubachians, 和 Laks 种族的没有血缘关系人, 对其与高海拔相关的基因进行了变异位点检测. 他们用第二代高通量测序分别测得了该 49 人与 2 组居住在海拔接近海平面的不同人种的与高海拔适应的 15 个基因的序列, 以及 20 个与海拔适应无关的区域. 他们总共得到了 1 066 个高质量的单核苷酸突变位点. 在 Laks 种族实验组中发现了一个 *HIF1A* 基

因中的非同义突变, 该突变已报到与提高氧气代谢有关; 在 Kubachians 组中发现了 13 个位于基因 *EGLN1* 的内含子保守区域的 SNPs. 这些变异说明了种族之间的差异性以及环境差异导致的基因差异性^[30].

3.3 CpG 位点测序

CpG 测序是指将 CpG 位点的 DNA 利用序列捕获技术进行捕获, 然后富集进行高通量测序的方法. CpG 位点的变异是导致疾病的重要原因之一. 至少有 1/3 的孟德尔遗传疾病是由于 CpG 位点的变异引起的^[31]. 例如, 在 50% 以上的人类癌症中都会变异的 *Tp53* 基因中, 大约有 30% 的变异都发生在 CpG 位点, 其中 5 个变异热点都在 CpG 位点中找到^[32]. 因此, 通过 CpG 位点测序并检测其变异位点对了解人类疾病意义重大, 近年来科学家们应用此方法陆续发现了很多位于 CpG 位点的变异.

Jin Billy Li 等用芯片获取了 21 号染色体上的 53 777 个 CpG 位点, 然后用 Illumina Genome Analyzer 进行测序, 经过分析发现了在这些位点中有 362 个新的 SNPs. 同时, 他们发现了在 CpG 位点中基因的转换频率为非 CpG 位点的 13.7 倍, 与之前科学家们研究发现的 CpG 位点的转换率为其它位点的 14~15 倍, 颠换频率为其它位点的 3~4 倍相一致^[33-35]. 他们的发现与染色体上甲基化程度的差异会导致人类基因组差异的假说相一致, 同时也表明了 CpG 位点目标测序是一种有效的检测一般和稀有变异的有效方法^[36].

4 转录组测序

转录组测序, 又称为 RNA-seq, 是指从总的 RNA 中提取出单链 RNA, 然后经过逆转录得到双链 cDNA, 然后对 DNA 进行测序的技术. 在检测人类基因组变异这个领域, 研究人员应用此种方法能够全面、精确地得到基因融合的信息. 虽然 DNA 测序也能够检测到基因融合, 但是转录组测序能够检测到更加稀有的融合. 同时, 对于 DNA 测序所解决不了的反式剪接^[37]以及邻近基因之间由于读码时未读到终止子, 而造成的相邻基因的融合^[38], RNA-seq 都能很好的解决. RNA 双端测序法, 已被广泛地应用于检测基因融合, 学者们运用这种方法, 发现了很多与肿瘤相关的基因融合, 这为治疗提供了更多潜在的靶点.

Kevin CH Ha 等用 Illumina RNA 测序法测了

5个 *BRCA1* 变异的乳腺癌细胞系(HCC1937、SUM149PT、SUM131502、HCC3153、HCC2337 细胞系)、3个 *BRCA1* 变异的原位乳腺癌组织、2个分泌型乳腺导管原位癌组织以及一个没有癌变的乳腺上皮细胞系的 RNA 序列。他们先后用单端、双端测序法进行了测序,发现有两个已知的基因融合通过这两种方法都能找到,同时还发现了3个新的基因融合,其中在细胞系 HCC3153 中的 *WWC1-ADRBK2* 基因融合以及在 *BRCA1* 变异的原位癌症组织中的 *ADNP-C20orf132* 基因融合都被 Sanger 测序法验证为真实的,这3个基因融合都只在自身的样本中出现。他们的发现表明,基因的融合在 *BRCA1* 变异的乳腺癌中出现的频率和重复性都很低,但是这种个体特异的基因融合也许在疾病的诊断和治疗中存在着潜在的价值^[39]。

Dan R. Robinson 等用 RNA 测序的方法测了一组乳腺癌细胞系以及组织的转录组序列。他们发现,每个癌症基因组中都有自己特定的基因融合,在这些融合中,他们发现了两组常发生的融合,它们分别与微管相关丝氨酸苏氨酸激酶(microtubule associated serine-threonine kinase, MAST)和 *Notch* 基因家族相关,都在乳腺上皮细胞的表型中发挥着重要的作用。有 *Notch* 基因融合的癌症样本,对 *Notch* 信号传导有着独特的灵敏的抑制作用, *MAST1* 或者 *MAST2* 基因融合的高表达在体内和体外都有促进增殖的作用。他们的这些发现表明,经常性发生的基因融合在一些癌症中有着重要的作用,同时还表明,转录组的测序能够起到检测稀有的基因融合的作用^[40]。

5 结 语

第二代高通量测序技术是生命科学领域一项跨时代的技术革新。它使得人类基因组的测序费用大大的降低,人类基因组测序从此进入万美元基因组时代。如此低廉的价格促使了近年来人类基因组重测序的迅猛发展。研究人员运用此方法发现了很多人类基因组上的突变,其速度之快、范围之全面是以往任何技术都无法比拟的。然而,由于受到其测序长度、人类基因组的复杂、庞大等原因的限制,检测重复区域以及某些长度的变异仍然困扰着科学家们。基于参考基因组的组装以及提高测序的读长能够很好地解决上述的问题。随着第二代高通量技术的发展,全基因组重测序、目标基因组测序、转录组测序能够更好地应用

于检测人类基因组突变位点,研究人员可以根据自己的需要选择合适的测序方法。我们相信,第二代高通量测序技术在检测基因突变领域的应用会更加频繁,为将来的医疗趋势——个性化医疗的发展提供坚实的保障。

参考文献(References):

- [1] CHECK E. Human genome: patchwork people [J]. *Nature*, 2005, 437(7062): 1084-1086.
- [2] SOUSSI T, BEROUD C. Assessing *TP53* status in human tumours to evaluate clinical outcome[J]. *Nature Reviews Cancer*, 2001, 1(3): 233-240.
- [3] ABRAHAM S C, NOBUKAWA B, GIARDIELLO F M, *et al.* Fundic gland polyps in familial adenomatous polyposis: neoplasms with frequent somatic adenomatous polyposis coli gene alterations[J]. *American Journal of Pathology*, 2000, 157(3): 747-754.
- [4] STRATTON M R, CAMPBELL P J, FUTREAL P A. The cancer genome[J]. *Nature*, 2009, 458(7239): 719-724.
- [5] WANG D G, FAN J B, SIAO C J, *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome[J]. *Science*, 1998, 280(5366): 1077-1082.
- [6] WILLIAMS R, PEISAJOVICH S G, MILLER O J, *et al.* Amplification of complex gene libraries by emulsion PCR[J]. *Nature Methods*, 2006, 3(7): 545-550.
- [7] MARGULIES M, EGHOLM M, ALTMAN W E, *et al.* Genome sequencing in microfabricated high-density picolitre reactors[J]. *Nature*, 2005, 437(7057): 376-380.
- [8] NIELSEN R, PAUL J S, ALBRECHTSEN A, *et al.* Genotype and SNP calling from next-generation sequencing data[J]. *Nature Reviews Genetics*, 2011, 12(6): 443-451.
- [9] QUINLAN A R, STEWART D A, STROMBERG M P, *et al.* PyroBayes: an improved base caller for SNP discovery in pyrosequences[J]. *Nature Methods*, 2008, 5(2): 179-181.
- [10] WU H, IRIZARRY R A, BRAVO H C. Intensity normalization improves color calling in SOLiD sequencing[J]. *Nature Methods*, 2010, 7(5): 336-337.
- [11] KIRCHER M, STENZEL U, KELSO J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies[J]. *Genome Biology*, 2009, 10(8): R83.
- [12] LI H, RUAN J, DURBIN R. Mapping short DNA sequencing reads and calling variants using mapping quality scores [J]. *Genome Research*, 2008, 18(11): 1851-1858.
- [13] LUNTER G, GOODSON M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads [J]. *Genome Research*, 2011, 21(6): 936-939.
- [14] LANGMEAD B, TRAPNELL C, POP M, *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. *Genome Biology*, 2009, 10(3): R25.
- [15] LI R, YU C, LI Y, *et al.* SOAP2: an improved ultrafast tool for short read alignment[J]. *Bioinformatics*, 2009, 25(15): 1966-1967.
- [16] LI H, DURBIN R. Fast and accurate short read alignment with Burrows-wheeler transform[J]. *Bioinformatics*, 2009, 25(14): 1754-1760.
- [17] BAINBRIDGE M N, WANG M, BURGESS D L, *et al.* Whole exome capture in solution with 3 Gbp of data[J]. *Genome Biology*, 2010, 11(6): R62.

- [18] LEY T J, MARDIS E R, DING L, *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome[J]. *Nature*, 2008, 456(7218): 66-72.
- [19] MARDIS E R, DING L, DOOLING D J, *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome[J]. *New England Journal of Medicine*, 2009, 361(11): 1058-1066.
- [20] PLEASANCE E D, CHEETHAM R K, STEPHENS P J, *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome[J]. *Nature*, 2010, 463(7278): 191-196.
- [21] AHN S M, KIM T H, LEE S, *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group[J]. *Genome Research*, 2009, 19(9): 1622-1629.
- [22] HILLMER A M, YAO F, INAKI K, *et al.* Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes[J]. *Genome Research*, 2011, 21(5): 665-675.
- [23] MAMANOVA L, COFFEY A J, SCOTT C E, *et al.* Target-enrichment strategies for next-generation sequencing[J]. *Nature Methods*, 2010, 7(2): 111-118.
- [24] LALONDE E, ALBRECHT S, HA K C, *et al.* Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing [J]. *Human Mutation*, 2010, 31(8): 918-923.
- [25] KELLER A, HARZ C, MATZAS M, *et al.* Identification of novel SNPs in glioblastoma using targeted resequencing[J]. *PLoS One*, 2011, 6(6): e18158.
- [26] BABENKO A P, POLAK M, CAVE H, *et al.* Activating mutations in the ABCC8 gene in neonatal diabetes mellitus[J]. *New England Journal of Medicine*, 2006, 355(5): 456-466.
- [27] PEARSON E R, FLECHTNER I, NJOLSTAD P R, *et al.* Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6. 2 mutations[J]. *New England Journal of Medicine*, 2006, 355(5): 467-477.
- [28] GLOYN A L, PEARSON E R, ANTCLIFF J F, *et al.* Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6. 2 and permanent neonatal diabetes[J]. *New England Journal of Medicine*, 2004, 350(18): 1838-1849.
- [29] BONNEFOND A, DURAND E, SAND O, *et al.* Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome[J]. *PLoS One*, 2010, 5(10): e13630.
- [30] LIM B C, LEE S, SHIN J Y, *et al.* Genetic diagnosis of Duchenne and Becker muscular dystrophy using next-generation sequencing technology: comprehensive mutational search in a single platform[J]. *Journal of Medical Genetics*, 2011, 48(11): 731-736.
- [31] COOPER D N, YOUSOUFIAN H. The CpG dinucleotide and human genetic disease[J]. *Human Genetics*, 1988, 78(2): 151-155.
- [32] OLIVIER M, EELES R, HOLLSTEIN M, *et al.* The IARC TP53 database: new online mutation analysis and recommendations to users[J]. *Human Mutation*, 2002, 19(6): 607-614.
- [33] KONDRASHOV A S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases[J]. *Human Mutation*, 2003, 21(1): 12-27.
- [34] HWANG D G, GREEN P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution[J]. *Proceedings of the National Academy of Sciences USA*, 2004, 101(39): 13994-14001.
- [35] SCHMIDT S, GERASIMOVA A, KONDRASHOV F A, *et al.* Hypermutable non-synonymous sites are under stronger negative selection[J]. *PLoS Genetics*, 2008, 4(11): e1000281.
- [36] LI J B, GAO Y, AACH J, *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations[J]. *Genome Research*, 2009, 19(9): 1606-1615.
- [37] LI H, WANG J, MOR G, *et al.* A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells[J]. *Science*, 2008, 321(5894): 1357-1361.
- [38] MAHER C A, PALANISAMY N, BRENNER J C, *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing[J]. *Proceedings of the National Academy of Sciences USA*, 2009, 106(30): 12353-12358.
- [39] HA K C, LALONDE E, LI L, *et al.* Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines[J]. *BMC Medical Genomics*, 2011, (4): 75.
- [40] ROBINSON D R, KALYANA-SUNDARAM S, WU Y M, *et al.* Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer[J]. *Nature Medicine*, 2011, 17(12): 1646-1651.