

To what extent are genetic variations and personal health linked?

遗传变异与人类健康

黄辉^①, 邓建莲^{①②}, 张欣鑫^{①②}, 李扬^①, 彭智宇^{①*}^① 深圳华大基因研究院, 深圳 518083;^② 中国科学院大学华大教育中心, 深圳 518083

* 联系人, E-mail: pengzhiyu@genomics.cn

2016-07-14 收稿, 2016-07-28 修回, 2016-08-01 接受, 2016-08-16 网络版发表

摘要 人类遗传变异常常是与生物学功能相关的, 遗传变异可影响人类健康. 不同种类的遗传变异往往对人类健康产生不同的影响, 深入了解不同变异对人类健康影响的机制, 将大大促进人类健康的研究. 而随着技术的进步, 尤其是高通量测序技术的广泛应用及相关技术的出现和发展, 必将极大完善基因诊疗方案. 本文侧重于遗传学变异对人类健康的影响, 探讨遗传变异和人体健康的直接联系, 以了解在人类健康中基因与环境、基因与基因之间的相互作用, 促进人们对人类健康的认识, 而这也是目前亟待探索的领域和话题.

关键词 遗传变异, 环境, 影响, 人类健康

阐明遗传变异在人类健康和疾病中的重要意义是21世纪的主要科学挑战之一^[1]. 为了研究遗传变异如何影响人类健康, 科学家们已经发起了大规模的致力于个体之间差异核苷酸的鉴定和分类的研究^[2,3]. 最初的研究主要集中于对单核苷酸多态性(single nucleotide polymorphisms, SNPs)的模式和频率范围的理解上^[4,5]. 随着结构变异对人类疾病患病率的贡献得到认识^[6,7], 现已成立了相关联盟并进行了系统的研究, 用以提高我国对该类变异的认识水平^[8~10].

2007年, J. Craig Venter的完整基因组序列作为首例个人基因组序列被发表^[2]. 随后不久, 又发表了第2例个人基因组序列, 即James D. Watson的基因组序列^[11]. 接着, 又对2个匿名个人的基因组进行了测序: 其中一个是中国汉族人^[12], 另一个是非洲人^[13]. 总的来说, 这些在人类基因组计划(human genome project, HGP)之后发表的研究, 迅速增加了我们对各种形式的人类遗传变异及它们的进化史和相互关系的认识. 然而, 我们对整个基因组上变异的位置和频率的理解仍然有限, 确定变异的分类是亟需解决的

问题.

1 人类存在多少遗传变异?

人类基因组含有约30亿个DNA碱基对, 除了同卵双胞胎之外, 没有两个人的基因型是完全相同的. 人与人之间的遗传差异(生物个性)大约为0.1%——这意味着两个人的基因组之间大约每1000个碱基就会有1个差异^[14].

人类基因组中最普遍的多态性(或遗传差异)是单碱基对差异, 即单核苷酸多态性. 比较2个不同的单倍体基因组, 平均每1000个碱基我们就可以发现1个SNP. 其他类型频率较低的多态性还包括拷贝数变异、插入、缺失、重复和重排等.

事实上, 研究结果一致表明, 人类的不同群体间, 有85%的遗传变异为群体特有, 而有15%的遗传变异为群体间共有^[14].

在人类健康和疾病与遗传变异的关联研究中, DNA测序新技术的应用为我们提供了新的视角, 同时也提出了新的问题.

引用格式: 黄辉, 邓建莲, 张欣鑫, 等. 遗传变异与人类健康. 科学通报, 2016, 61: 2772~2781

Huang H, Deng J L, Zhang X X, et al. Genetic variations and human health (in Chinese). Chin Sci Bull, 2016, 61: 2772~2781, doi: 10.1360/N972016-00586

本文讨论了不同类型的遗传变异对人类健康的影响,论述了遗传变异与人类健康之间相关作用机制,同时阐述了检测技术进步带来科研、临床应用上的进步,并具体说明遗传病的预防和治疗进展,为人类健康与相关研究提供参考。

2 人类遗传变异对健康的影响

研究表明,人类大量的表型变化是由遗传因素和非遗传因素(环境等)共同引起的。许多临床疾病都与基因突变有关。

人类遗传变异通常被分成常见型和罕见型,以人群中最小等位基因频率(minor allele frequency, MAF)来区分。常见变异又可称为多态性,是指群体中的最小等位基因频率至少为1%的遗传变异,而罕见变异的群体最小等位基因频率则小于1%。从广义上说,人类基因组的变异可以被分成两类:单核苷酸变异和结构变异。绝大多数的遗传变异是中性的,不会导致表型变异。

人类基因组变异有多种形式,例如单核苷酸多态性、可变数串联重复序列、转座因子的有无以及结构变异、染色体变异等。这些变异在基因组中普遍存在并且广泛分布,个体间表型差异和许多遗传性疾病及一些复杂疾病的易感性被认为是由这些变异造成的。

2.1 SNP与人类健康

个体间表型差异、许多遗传性疾病及一些复杂疾病的易感性与基因组的关联因素即基因组中的变异,其中最主要的就是单核苷酸多态性。

单个核苷酸的多态性是人类可遗传的变异中最常见的一种,占有已知多态性的90%以上。根据目前的测序结果大致估计人类基因组至少包含1100万个SNP,其中700万个最小等位基因频率大于5%^[15],其余在1%~5%之间。大多数的SNP信息被收录在单核苷酸多态性数据库(dbSNP)中。除了单核苷酸多态性外,还有不计其数的罕见和新发的单核苷酸变异,在某些案例中变异与疾病的共分离仅仅发生在一个核心家系或者单个个体中。SNP在人类基因组中广泛存在,平均每500~1000个碱基对中就有1个,估计其总数可达300万个甚至更多。大量存在的SNP位点,使人们有机会发现与各种疾病,包括肿瘤相关的基因组突变;从遗传基因组学方面来看,与慢性阻塞性

肺疾病(chronic obstructive pulmonary disease, COPD)相比,哮喘-慢性阻塞性肺疾病重叠综合征(asthma COPD overlap syndrome, ACOS)病人,在基因水平上由于单个核苷酸变异所引起的DNA序列多态性对ACOS患者已经起到了不可忽视的作用^[16]。从实验操作来看,通过SNP发现疾病相关基因突变要比通过家系来得容易;有些SNP并不直接导致疾病基因的表达,但由于它与某些疾病基因相邻,而成为重要的标记,用于寻找遗传疾病相关基因,对致病基因进行定位和克隆。SNP的高密度和遗传稳定性,使它们适用于许多研究疾病基因的策略,如家系的连锁分析、隔离群体的连锁不平衡分析、疾病与病因的关联分析、肿瘤的杂合性缺失研究等。随着DNA测序技术的纵深发展,人们愈来愈相信基因组中的这类SNP多态性与个体的表型差异、不同群体对疾病特别是复杂疾病的易感性以及对各种药物的疗效和对环境因子的反应等相关。

从对生物的遗传性状的影响上来看,位于蛋白质编码区的SNP称为编码单核苷酸多态(coding SNP, cSNP)。cSNP又可分为两种:一种是同义cSNP (synonymous cSNP),即SNP所致的编码序列的改变并不影响其所翻译的蛋白质的氨基酸序列,突变碱基与未突变碱基的含义相同;另一种是非同义cSNP (non-synonymous SNP),指碱基序列的改变可以使其为转录本翻译的蛋白质序列发生改变,从而影响了蛋白质的功能,这种改变常是导致生物性状改变的直接原因,cSNP中约有一半为非同义cSNP,cSNP又可分为错义突变、无义突变等。

多数孟德尔单基因遗传病是由基因组中蛋白编码区的碱基突变导致蛋白功能和表型改变所致。常见的基因突变包括碱基置换(错义突变、无义突变、终止密码突变)、密码子缺失或插入、移码突变、融合基因、拼接区突变、动态突变(主要指三联核苷酸重复)以及大片段缺失或插入等。基因突变产生的分子生物学效应包括功能缺失、功能增强、形成新特征以及显性负效应等。常见的单基因病有分子病,如镰形红细胞贫血症和地中海贫血等;还有生化代谢异常,如因氨基酸代谢异常所致的苯丙酮尿症,因糖代谢异常所致的半乳糖血症等。

目前对于SNP的疾病遗传学研究主要集中于基因内或与基因关联的SNP,用于寻找遗传疾病相关基因,对致病基因进行定位和克隆。SNP的高密度和

遗传稳定性,使它们适用于许多研究疾病基因的策略,如家系的连锁分析、隔离群体的连锁不平衡分析、疾病与病因的关联分析、肿瘤的杂合性缺失研究等。

2.2 结构变异与人类健康

基因组结构变异通常是指基因组内大于1 kb的DNA片段缺失、插入、重复、倒位、易位、转座子元件以及DNA拷贝数目变化(CNVs)等。随着人类基因组测序的广泛应用,使基因测序变得常规和普遍,现也有人将大于50 bp的基因组异常称为结构变异^[17]。人类基因组结构变异涉及数千片段不连续的基因组区域,含数百万DNA碱基对,可含数个基因及调控序列,多种基因功能因此缺失或改变,导致机体表型变化、疾病易感性改变或发生疾病。对基因组结构变异的研究,有助于用动态的观点全面分析基因组遗传变异得到整合的基因型,理解结构变异的潜在医学作用及机体整体功能的复杂性。

将人类基因组参考序列作为参照标准,个体基因组结构变异可定量描述为拷贝数获得(插入和重复)及拷贝数丢失(缺失);或是位置变异(易位和重排)及方向变异(倒位)。与SNP比较,结构变异虽然发生的频率较低,但累及的序列长度却大大超过了SNP,因此对人类健康和疾病的影响更为显著。

根据遗传物质是否增减,可以将基因组结构变异分为平衡重排和非平衡重排。平衡重排包括倒位和某些易位,没有遗传物质的丢失或获得,这类结构变异可以通过基因组序列比较或是细胞遗传学方法检测。具有临床表型的平衡重排常通过染色体断裂点检测而鉴定,没有临床表现的平衡重排称携带者。由于结构变异引起携带者基因组不稳定,携带者虽然没有临床表现,但携带的结构变异将对后代表型产生影响。据推测,人类基因组中可能存在20%的平衡重排结构变异。

结构变异或引起个体的病理表型,或仅仅在群体中表现为一种多态。有病理表型的结构变异在临床上表现为多种综合征和疾病,没有临床表现的变异能够引起携带者基因组不稳定,减数分裂中产生不等交换对后代表型产生影响^[18]。

结构变异由于覆盖的核苷酸总数大大超过SNP的总数,因此对个体表型的影响很大。结构变异中的基因拷贝数变化通过改变基因剂量、扰乱基因表达、

引起个体表型改变甚至产生疾病,如个体防御素基因拷贝数差异可引起个体抗感染能力差异,导致疾病产生就是典型的例子。一般说来,结构变异可通过几条途径影响个体的表型:一是直接影响基因的剂量(拷贝数变化)导致基因表达蛋白量改变;二是通过改变基因位置影响相关基因调节,间接影响基因表达诱发疾病;三是引起基因重排,诱发基因结构改变(基因断裂、基因融合)产生疾病基因。此外,结构变异的不同类型相互作用或结合SNP及环境因子共同对个体的表型产生影响。例如:Williams-Beuren综合征中发现89%的患者染色体7q11.23有1.55 Mb的缺失,含26~28个基因,同时发现患者在这个区域内一个高达25%的2 Mb臂内倒位,推测由于倒位序列存在,引起减数分裂异常配对,产生不等交换引发缺失产生,导致疾病发生^[19]。

对于大的变异(主要是获得或者缺失几百kb的片段)在人群中是十分少见的(<1%),但是会导致很大一部分疾病。对于多拷贝基因家族的基因拷贝数量的变异,个体健康的影响主要是增加个体对某些疾病的易感性,使个体更容易患有某些疾病^[16]。如自闭症患者基因组中存在的较高频率的CNVs及染色体16p11.2片段缺失与该病的易感性有关。

结构变异几乎同所有的遗传性疾病有关,从罕见疾病、遵循孟德尔遗传规律的单基因疾病、到感染性疾病、代谢性疾病、药物毒物代谢等。

2.3 染色体变异与人类健康

从更大的层面来看,众所周知,染色体数目和结构的异常也会导致疾病,这类疾病称为染色体病。染色体数目异常是临床上最主要的染色体病,其中以13, 18, 21号及X, Y染色体异常最为常见,占染色体非整倍体畸变95%以上。这些染色体异常占全部染色体病的80%~90%^[20]。

染色微结构异常疾病是由于染色体微小片段缺失或重复,使正常基因组剂量发生改变而导致临床可识别的一组疾病。染色体微结构异常是用常规的染色体显带方法不能或不容易发现的染色体结构异常。随着分子细胞遗传技术的发展,特别是最近两年CGH微阵列方法的引进,被发现的染色体微结构引起的疾病将会越来越多。如22q11.2微缺失综合征是由于22号染色体长臂近着丝粒端微片段22q11.21缺失引起的遗传综合征,其遗传病理基础是22q11.2

片段的微缺失. 主要临床表现为心脏畸形、面容异常、胸腺发育不良、腭裂及低钙血症5个症状, 又称DiGeorge综合征(DiGeorge syndrome, DGS)和腭心面综合征(velocardiofacial syndrome, VCFS). 近年研究认为22q11.2微缺失综合征与学习认知障碍、精神异常、发育迟缓等有关. 其发病率为活产新生儿的1/4000~1/3000, 是人类最常见的染色体微缺失综合征. 另外, 该区域的微重复异常也会导致疾病, 称为22q11微重复综合征, 临床表现与22q11.13缺失综合征相似, 但有其特殊表型, 其中包括上位眉毛, 眼眶下斜、轻度小颌、脸部细长等. 染色体微缺失微重复异常主要是由于缺失或重复的染色体区段包含有个体发育过程中发挥重要作用的关键基因, 缺失或重复导致基因功能异常从而致病. 相对来说, 染色体重复引起的遗传学效应比缺失为轻^[21].

2.4 遗传易感性与人类健康

总体上, 孟德尔遗传模式的疾病往往表现为相对不常见或发病率低, 且发病早, 这些基因中某些突变常有高外显率——这意味着, 携带疾病易感基因型的个体最终发展成疾病的概率相对较高(大于50%).

相对于前者, 多基因遗传病(polygenic disease)是指累加基因和环境因素共同影响而导致的疾病, 因此, 这种遗传方式又称多因子遗传(multifactorial inheritance). 疾病常出现家族倾向, 但不表现出孟德尔遗传规律. 常见的成人疾病如高血压、冠心病、痛风、糖尿病、精神分裂症及抑郁症等, 都是多个基因和环境因素共同作用的结果, 而且遗传基础复杂, 故也把这些疾病称为复杂遗传病(complex disease). 在多基因遗传病中, 若干作用微小但有累积效应的致病基因构成了个体患某种疾病的遗传因素, 这种由遗传基础决定一个个体患病风险称为易感性(susceptibility). 而由遗传因素与环境因素共同作用并决定一个个体是否易于患病的可能性称为易患性(liability)^[21].

环境因素对不同个体产生的影响是不同的, 并且环境和遗传异质性的综合影响导致了病因异质性. 病因异质性是指一种疾病普遍发生在整个群体中, 多组病例(例如乳腺癌簇)表现出相似的临床特征, 但实际上具有不同病因的现象. 通过病因学异质性的疾病案例的发现和检测, 使我们能更加深入地研究特定疾病的病因学, 寻找可能的病因. 这些研究的结

果也可能会凸显出基因-基因之间相互作用及基因-环境之间相互作用在疾病进程中的重要性. 运用分子流行病学技术确定病因异质性对于疾病的分析而言是一个十分重要的步骤, 以此促进疾病预防工作^[22].

许多常见病, 如心血管病、糖尿病、肿瘤等, 属于多基因病, 其患病风险有一部分取决于遗传学因素, 另外一部分则取决于环境因素. 例如吸烟、饮食和锻炼. 而且它们的遗传学因素来自于多个基因的共同作用, 每个基因对于患病风险的贡献不足5%, 所以很难用单个基因进行诊断. 一般认为这类疾病不具有遗传性, 但是通过对越来越多的案例研究发现这一类疾病具有家族聚集性, 不排除有一定的遗传性因素. 常见的大类疾病多有此类模式, 如癌症、心脏疾病、自身免疫疾病、高血压、糖尿病、肥胖症等. 尽管多基因疾病并不认为完全由遗传因素导致, 或者说由父母遗传给后代, 但是家族史是疾病发生的其中一个风险因素, 此外还有一定的环境因素起作用.

多基因疾病患病群体数量巨大. 在1996年, 美国就有8511845个自身免疫疾病患者^[23]; 根据美国国家心肺血液研究所的统计结果, 大约有5000万美国人患有高血压, 平均每1000人就有217人患病(美国国家卫生情报局); 糖尿病的发病率同样惊人, 1000万美国人被确认为糖尿病患者, 发病率为65/1000, 至少还有1600万有糖尿病的倾向而未被确诊(美国国家卫生情报局), 且每年都有新增的案例为80万(美国疾病防治中心); 美国是个肥胖症高发的地区, 大约有4000万美国成年人患有肥胖, 过重人群占比达到了57%(美国疾病防治中心); 此外也有2200万美国人存在心脏疾病(美国疾病防治中心). 根据这些参考的数据, 我们可以看到这类多基因疾病的影响广泛. 目前对于多基因疾病的研究仍处于基础起步阶段, 测序技术的发展使得我们可以获取大量的基因组信息, 但是多基因疾病的具体基因模式尚未明确, 后续还需要进行深入的研究.

由于人类遗传的多态性, 不同个体对环境致病因素的易感性也有差异. 针对与环境中物理、化学或生物因素发生相互作用的蛋白的编码基因, 识别其基因组多样性和结构与功能的关系, 将有助于发现特定环境因子致病的人群, 并制定相应的预防措施和环境保护策略.

2.5 环境因素和遗传变异的相互作用与人类健康

几乎所有的人类疾病都是由遗传易感性因素和可变的环境因素相互作用的结果,环境因素广义的定义包括感染、化学、物理、营养、行为等因素,这也许是了解遗传学和环境在疾病发展中作用的最重要的事实.许多人倾向于将疾病的原因归类为遗传或环境.事实上,一些罕见疾病,如亨廷顿或Tay-Sachs病,可能是缺乏单一基因产物的结果,但这些疾病在人类所有的疾病中只占一个非常小的比例.常见的疾病,如糖尿病或癌症,是由于遗传和环境因素的复杂相互作用的结果.

几乎所有的疾病都与基因结构的变化有关,甚至所谓的单基因疾病实际上也是由遗传和环境因素的相互作用而发展的.例如,苯丙酮尿症(PKU)是由于苯丙氨酸(PA)代谢途径中的酶缺陷,使得苯丙氨酸不能转变成为酪氨酸,导致苯丙氨酸及其酮酸蓄积,并从尿中大量排出.在正常的蛋白质摄入的情况下,苯丙氨酸积累且产生神经毒性.当遗传变异(苯丙氨酸代谢途径中的酶缺陷)和环境暴露(膳食苯丙氨酸)共同存在时PKU才会发生.

遗传变异不会引起疾病,而是影响一个人对环境因素的易感性,我们不能继承一种疾病状态,相反,我们继承了一组易感因素,如环境因素的某些影响,因此继承了特定疾病的高患病风险.这一概念也解释了为什么相同的环境因素会对个体产生不同的影响.例如,一些在胆固醇水平正常范围内的健康个体却在40岁时患心肌梗死.其他人尽管吸烟、饮食不规律和肥胖,却似乎对心脏病免疫.这就是相同的环境因素的遗传变异差异造成的.

我们都携带遗传变异,这些遗传变异增加了我们对某些疾病的易感性.通过识别和表征基因与环境的相互作用,我们有更多有效的方法进行目标干预.许多疾病的遗传危险因素还没有被确定,并且其他基因和基因与环境因素的复杂相互作用还没有被理解.临床和流行病学有关进一步描述这些因素和它们相互作用的研究是必要的.然而,随着我们对遗传变异和环境因素的理解和认识更加深入,将有利于具有疾病高患病风险的个人目标干预策略的制定.

烟草烟雾中含有大量可能导致基因损伤的化学致癌物.一些基因修复通路可以修复这种损伤,而我們也可通过对通路中基因的研究来了解为什么有些吸烟者患肺癌,其他人却不患,从而确定主要生物学

候选基因.在一项研究中,科研人员对负责基因修复的两个基因变异与吸烟频率和肺癌伴发情况的相关性进行了检测^[24].在1091例肺癌患者和1240名正常对照检测XRCCI基因的一个功能性突变及ERCC2基因中两个功能性突变,并将病例组和对照组分为重度吸烟者与不吸烟者.研究人员发现带有XRCCI基因突变的不吸烟者比正常基因型的不吸烟者患肺癌的风险大2.4倍.带有XRCCI突变的重度吸烟者与其带有正常基因型的对照组相比,在患肺癌风险上降低50%.当这2个基因的3个突变同时存在于极端基因型个体(基因中带有5或6个变异)中,对这一个体进行检测,与没有变异的个体相比,非吸烟者的肺癌风险大5.2倍且在重度吸烟者中风险降低70%.在重度吸烟者中的这些遗传变异的保护作用可能是由吸烟引起这些保护基因活性差异性增加引起的.类似的基因相互作用也在其他基因中发现,如ERCC1基因.这些研究说明了识别与人类行为相关疾病遗传变异的重要性.为充分了解基因易感性及环境因素是如何影响公众健康的,需要对多个突变的联合影响进行更多的研究.生物信息学的进步在这一领域的研究中起到了关键的作用.转录调控区突变SNP数据库可以用于多态性调控元件的功能验证,如在编码阶段II解毒酶和许多基因启动子区域发现的抗氧化调节元件^[25].比较序列分析方法对人类遗传研究也正变得越来越有价值,因为它们提供了一种对单核苷酸多态性在蛋白质功能或基因调控中潜在的有害影响进行排序的手段^[26].运用非同义单核苷酸多态性的大规模分析方法来预测一个特定的突变是否会对蛋白质功能产生影响^[27],有利于在疾病病因学研究中单核苷酸多态性的筛选,也可用于被认为与不同个体对环境因素(如饮食,压力和社会经济因素)的不同反应相关的突变的功能分析.

3 检测技术进步带来科研、临床应用上的进步

近几年,随着基因测序技术的进步,特别是二代测序(next-generation sequencing, NGS)技术的出现和发展,测序效率得到了显著提升,测序时间明显缩短,测序费用大大降低,基因检测技术逐渐成为科学研究和临床诊断的热点,得到了飞速的发展,大量科研和临床应用上的成果不断涌现.2009年3月,约翰

霍普金斯大学的科研人员运用NGS外显子测序技术发现遗传性胰腺癌的一个新的致病基因PALB2, 研究成果发表在*Science*杂志上^[28]。2009年9月, 华盛顿大学的科研人员将其利用NGS技术研究罕见弗里曼谢尔登综合征发现MYH3致病基因突变的成果发表在*Nature*杂志上^[29]。之后, 科研人员利用NGS技术, 不断发现与遗传相关的致病基因, NGS技术的运用大大加速了科学研究。

NGS主要包括全基因组测序(whole-genome sequencing, WGS)、全外显子测序(whole-exome sequencing, WES)和目标区域测序(targeted regions sequencing, TRS)。

全基因组测序是对未知基因组序列的物种进行个体的基因组测序。因为人类全基因组测序已经完成, 现在临床和科研上所进行的为全基因组重测序(whole genome re-sequencing)。全基因组重测序是对已知基因组序列的物种进行不同个体的基因组测序, 并在此基础上对个体或群体进行差异性分析。其测序覆盖面广, 能检测个体基因组中的全部遗传信息; 准确性高, 其准确率可高达99.99%。但由于人类基因组过于庞大, 往往需要通过重复测序或双端测序, 测序成本由此大幅度提高, 检测费用因此高昂, 对于一般临床诊断和普通科研项目难以承担其花费。

全外显子测序是指利用序列捕获技术将全基因组外显子区域DNA捕捉并富集后进行高通量测序的基因组分析方法。到目前为止, 利用全外显子测序在科研及临床应用中已经取得了巨大的成果, 对于致病基因的鉴定起到了显著的推动作用。而这些检测已经不单单局限在单基因遗传疾病的鉴定上, 在多基因影响的复杂疾病中也发现了大量的相关基因。在多种遗传性疾病的致病基因的鉴定中已逐步完全确定病因, 如进行性肌营养不良、白化病、成果不全和无丙球蛋白血症等发现了致病突变。另外, 在遗传性肿瘤的鉴定上也取得了突出的进展, 如小细胞肺癌、慢性淋巴细胞性白血病等。

目标区域测序是目前常用的基因芯片技术, 基于DNA杂交原理, 通过定制目标基因组区域的探针, 与基因组DNA进行杂交, 将目标区域DNA富集后进行NGS测序的技术手段。基因芯片技术对于已知基因突变的筛查具有明显优势, 可以快速、全面地检测出目标基因突变。2010年剑桥大学的科研人员利用基因分型芯片联合连锁分析技术, 发现小头畸形的新

基因WDR62, 其研究成果发表在*Nat Genet*上^[30]。通过利用大量样本对目标区域进行研究, 有助于发现和验证疾病相关候选基因或相关位点, 在科研和临床应用上具有很大的前景。

目前, 利用NGS技术进行的基因检测已经大大扩展了其范围, 从单一遗传病的筛查、检测到复杂病和遗传性肿瘤的检测, 现已广泛应用在无创产前检测、新生儿基因检测、遗传病家系分析和检测及携带者基因检测上, 同时对药物基因组学和营养基因组学等领域也有巨大的推动作用。基因检测技术表现出极大的科研和临床应用价值。

4 遗传病的预防和治疗进展

遗传性疾病大多存在治疗困难、预后差等难题, 减少出生缺陷, 关键在于预防。目前可通过“三道防线”预防出生缺陷: 一级预防——婚前家族史咨询和孕前优生筛查; 二级预防——产前筛查和产前诊断; 三级预防——新生儿疾病筛查。从做好孕前咨询准备、充分孕前筛查, 接受规范的产前筛查和产前诊断, 同时配合新生儿疾病筛查等多方面入手, 最大程度地降低出生缺陷的发生。目前基因检测技术可与三道防线有机的结合, 极大程度地促进了遗传性疾病的诊断, 大大降低了出生缺陷率。

近30年来人类对自身生殖过程有了深入的研究, 同时发展起来的辅助生殖技术和分子遗传学技术通过结合, 使人们可以在植入前的早期胚胎进行遗传性疾病检测, 从而筛选出正常的胚胎进行宫腔移植, 即为植入前遗传性诊断。与产前诊断不同的是, 植入前诊断是在妊娠之前进行的, 因此避免了选择性流产, 减少对孕妇个人的伤害。一般来说, 可进行植入前遗传性诊断的疾病主要为单基因遗传病、三联体重复序列异常疾病和染色体异常。

随着基因编辑技术在临床上的成功应用, 预示着基因编辑治疗时代的来临, 给遗传性疾病患者治愈疾病带来曙光。基因编辑技术及基因编辑的三大利器: ZFN(锌指核酸酶)、TALEN(转录激活样效应因子核酸酶)和CRISPR/Cas9(成簇规律间隔短回文重复技术)。

基因编辑技术是指能够让人类对目标基因进行“编辑”, 可以对基因组完成精确修饰的一种技术, 可完成基因定点InDel突变、敲入、多位点同时突变和小片段的删除等, 可在基因组水平上进行精确的基

因编辑. 在科研领域, 该技术可以快速构建模式动物, 节约大量科研时间和经费; 在医疗领域, 基因编辑技术可以更加准确、深入地了解疾病发病机理和探究基因功能, 可以改造人的基因, 达到基因治疗的目的等. 因此, 基因组编辑具有极其广泛的发展前景和应用价值.

2014年3月, 宾夕法尼亚大学的科研人员在 *N Engl J Med* 上发表了利用锌指核酸酶 ZFN 技术基因编辑细胞治疗 12 名艾滋病患者^[31], 这是第一次成功使用基因治疗的报道. 2015年11月, Waseem Qasim 对外宣布, 他们利用 TALEN 基因编辑技术改造的 UCART19 细胞株, 成功缓解了 Layla 的急性淋巴细胞性白血病^[32], 这是第二次基因编辑技术在临床上的成功应用. 而 CRISPR/Cas9 技术自问世以来, 就有着其他基因编辑技术无可比拟的优势, CRISPR/Cas9 系统是一个天然存在的原核生物 RNA 干扰系统, 其介导的基因组编辑是由 crRNA 指导的, 对靶序列的识别是 RNA 与 DNA 的碱基配对过程, 相比蛋白质对 DNA 序列的识别要精确更多, 降低了脱靶切割的几率, 减低了细胞毒性. 而且 CRISPR/Cas9 的构建仅仅需要设计与靶序列互补的 RNA 即可, 过程相对于 TALEN 更为简单和廉价, 普通的实验室也可以自行完成构建, 这大大提高了基因操作的效率及简便性. 但是, CRISPR/Cas9 系统也存在着一些不足. 首先, Cas9 蛋白对于目标序列的切割不仅仅依靠 crRNA 序列的匹配, 在目标序列附近必须存在一些小的前间区序列邻近基序(PAM), 如果目标序列周围不存在 PAM 或者无法严格配对, 则 Cas9 蛋白不能对任意序列进行切割. 其次, 和 ZFN 及 TALEN 技术一样, CRISPR/Cas9 也面临着如何控制双链断裂之后的非同源末端连接修复可能随机产生细胞毒性的问题. 然而技术的发展有助于解决这些问题, 2015年1月, *Nature* 发表了标题为“Promoterless gene targeting without nucleases ameliorates haemophilia B in mice”的研究论文^[33], 发布了一种超越 CRISPR 的基因组编辑新技术, 这一研究发现可能成为 CRISPR 基因编辑技术的一种替代选择, 从而使多种疾病有望得到治愈. 更加振奋人心的是, 2016年5月2日, *Nat Biotechnol* 在线发表了韩春雨的研究成果“DNA-guided genome editing using the *Natronobacterium gregoryi* Argonaute”^[34]. 该研究成果找到了对基因组位点编辑范围更广的基因编辑工具. 该工具完全不同于以

RNA 为向导的 CRISPR/Cas9 基因编辑技术: 这种从古细菌来源的 Argonaute (简称 NgAgo), 利用短链 DNA 作向导, 真正实现了对基因组的任意位置进行切割, 将基因编辑的可能性推入了更广泛的境地.

5 总结与展望

人类基因组作为一个十分稳定却又存在各种变异的体系, 对其变异或多态性的研究正日益显示其重要性, 因为后者有助于基因的鉴别和定位, 同时通过建立变异与表型或疾病(如各种心血管病及癌症)风险之间的关系, 可以对人类进化、种群多样性及复杂疾病的诊断和治疗产生巨大的推动作用.

随着 DNA 测序技术的发展和运用, 大大促进了人类对遗传变异与表型差异关系的研究和不同群体中的疾病特别是复杂疾病的遗传易感基因的鉴定; 也促进了遗传变异与各种药物的耐受性和对环境因子的反应的相关性研究.

几乎所有的人类遗传变异的生物学意义都是微不足道的, 也就是说, 这些遗传变异并没有适应意义. 某些变异如中性突变(neutral mutation)改变了所表达蛋白质的氨基酸序列, 但并不会改变蛋白质的功能; 另外一些变异如沉默突变(silent mutation)则不改变氨基酸序列. 此外, 人类基因组的 DNA 序列中只有一小部分为编码序列(最终可翻译成蛋白质)或调节序列(可以影响基因表达的水平、时效性和组织特异性的序列), 在 DNA 序列的编码序列和调节序列之外的区域发生的变异绝大多数并没有已知的功能影响. 然而, 一些遗传变异具有正向选择作用, 这让个体能适应不断变化的环境.

几乎所有细胞都携带含有大量遗传变异的基因组信息, 这些包含突变的基因组信息作为转录翻译的模板, 形成了个体蛋白质行使生物学功能的特定机制. 生物学家普遍认为基因组上的遗传变异可影响生物的多个层面, 包括细胞、生物化学、生理和药物等. 目前许多研究集中于探索遗传变异和疾病风险的关系. 现已逐渐认识到常见疾病如心脑血管疾病、高血压、癌症、糖尿病和许多精神疾病中, 疾病进程的不同方面可受不同基因和不同遗传变异的影响. 例如, 某些基因的遗传变异可与疾病发生时的易感性相关, 而另外一些基因或遗传变异则可能与疾病进程相关, 其影响可至个体死亡; 此外, 又有另外一套完全不同的基因集可能与个体对药物的反应相

关; 同样, 也可能有另外一套基因集与个体对于疾病相关的环境风险因子的反应程度相关。最终, 不同的基因可使个体易患某种特定的疾病, 进而对其寿命产生影响。

关于遗传变异是如何增加疾病的易感性, 我们的研究仍处于起步阶段。许多人认为, 若不考虑环境因素, 我们将无法全面地了解遗传变异在人类健康中的作用。由于对基因组、转录组、蛋白组和代谢组的生物学变异的研究技术的突破, 使我们对基因在疾病中的作用的观念框架得到很大转变。多个基因同时发挥作用, 并结合着个体的年龄、性别、饮食和身体情况等情况修饰影响着个体对疾病的易感性。这便是一个特定疾病易感性的整体框架。在大部分情况下, 我们仍然处在记录疾病复杂性、寻找遗传易感性基因、了解疾病异质性和建立疾病风险模型的阶段, 其中的疾病风险模型的建立会基于我们对于生

物学的理解而进行不断的修正。

总的来说, 20世纪对人类疾病与健康的研究集中于鉴定影响个体遗传易感性的基因和环境因素。在这期间, 我们对造成严重临床后果的孟德尔疾病的研究已经有了很大的进步。然而, 孟德尔遗传模式无法明确地解释常见慢性疾病的遗传易感性, 对于这些常见的慢性疾病, 研究人员根据其明显的家族聚集性, 倾向认为这类疾病也具有一定的遗传性。同样, 对环境和社会流行病学的研究在解释许多环境因素如饮食、运动和压力与疾病风险的相关性也获得了巨大的成功。然而, 这些环境因素仍然不能完整的解释为什么在不同的人群中某些疾病发病率存在差异。目前研究认为遗传与环境的相互作用可解释大部分人群中大部分疾病的发生, 通过对遗传因素和社会环境因素和的整合研究, 复杂疾病的发病机制有望得到进一步阐明。

参考文献

- 1 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001, 409: 860–921
- 2 Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. *Science*, 2001, 291: 1304–1351
- 3 International HapMap C. A haplotype map of the human genome. *Nature*, 2005, 437: 1299–1320
- 4 Frazer K A, Ballinger D G, Cox D R, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 2007, 449: 851–861
- 5 Hinds D A, Stuve L L, Nilsen G B, et al. Whole-genome patterns of common DNA variation in three human populations. *Science*, 2005, 307: 1072–1079
- 6 Iafrate A J, Feuk L, Rivera M N, et al. Detection of large-scale variation in the human genome. *Nat Genet*, 2004, 36: 949–951
- 7 Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*, 2004, 305: 525–528
- 8 Eichler E E, Nickerson D A, Altshuler D, et al. Completing the map of human genetic variation. *Nature*, 2007, 447: 161–165
- 9 Kidd J M, Cooper G M, Donahue W F, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 2008, 453: 56–64
- 10 Redon R, Ishikawa S, Fitch K R, et al. Global variation in copy number in the human genome. *Nature*, 2006, 444: 444–454
- 11 Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 2008, 452: 872–876
- 12 Wang J, Wang W, Li R, et al. The diploid genome sequence of an Asian individual. *Nature*, 2008, 456: 60–65
- 13 Bentley D R, Balasubramanian S, Swerdlow H P, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 2008, 456: 53–59
- 14 National Institutes of Health. NIH Curriculum Supplement Series. Bethesda (MD): National Institutes of Health (US), 2007. <http://www.ncbi.nlm.nih.gov/books/NBK20363/>
- 15 Kruglyak L, Nickerson D A. Variation is the spice of life. *Nat Genet*, 2001, 27: 234–235
- 16 Xu F, Dong J C. Clinical research progress of Asthma—COPD overlap syndrome (in Chinese). *Chin Gen Prac*, 2016, 19: 500–506 [徐飞, 董竞成. 哮喘—慢性阻塞性肺疾病重叠综合征的临床研究进展. *中国全科医学*, 2016, 19: 500–506]
- 17 Alkan C, Coe B P, Eichler E E. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011, 12: 363–376
- 18 He Y S, Zhang W, Yang Z Q. Structural variation in the human genome (in Chinese). *Hereditas(Beijing)*, 2009, 31: 771–778 [何永蜀, 张闻, 杨照青. 人类基因组结构变异. *遗传*, 2009, 31: 771–778]
- 19 Cuscó I, Corominas R, Bayés M, et al. Copy number variation at the 7q11. 23 segmental duplications is a susceptibility factor for the Williams-Beuren syndrome deletion. *Genome Res*, 2008, 18: 683–694

- 20 Huang Y Y. Prenatal diagnosis of chromosomal diseases (in Chinese). *Chin J Clinicians (Elect Ed.)*, 2012, 6: 2853–2856 [黄艳仪. 染色体疾病的产前诊断. 中华临床医师杂志(电子版), 2012, 6: 2853–2856]
- 21 Lu G H, Xu X M. *Clinical Genetic Counselling* (in Chinese). Beijing: Peking University Medicine Press, 2007. 91 [陆国辉, 徐湘民. 临床遗传咨询. 北京: 北京大学医学出版社, 2007. 91]
- 22 Rebbeck T R, Walker A H, Phelan C M, et al. Defining etiologic heterogeneity in breast cancer using genetic biomarkers. *Prog Clin Biol Res*, 1996, 396: 53–61
- 23 Rose N R, Mackay I R. *The Autoimmune Diseases*. 3rd ed. St. Louis, MO: Academic Press, 1998
- 24 Zhou W, Liu G, Miller D P, et al. Polymorphisms in the DNA repair genes *XRCC1* and *ERCC2*, smoking, and lung cancer risk. *Cancer Epidemiol Biomarkers Preven*, 2003, 12: 359–365
- 25 Miller D P, Liu G, De Vivo I, et al. Combinations of the variant genotypes of *GSTP1*, *GSTM1*, and *p53* are associated with an increased lung cancer risk. *Cancer Res*, 2002, 62: 2819–2823
- 26 Wang Z, Fan H, Yang H H, et al. Comparative sequence analysis of imprinted genes between human and mouse to reveal imprinting signatures. *Genomics*, 2004, 83: 395–401
- 27 Clifford R J, Edmonson M N, Nguyen C, et al. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, 2004, 20: 1006–1014
- 28 Jones S, Hruban R H, Kamiyama M, et al. Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science*, 2009, 324: 217–217
- 29 Ng S B, Turner E H, Robertson P D, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 2009, 461: 272–276
- 30 Nicholas A K, Khurshid M, Désir J, et al. *WDR62* is associated with the spindle pole and is mutated in human microcephaly. *Nat Genet*, 2010, 42: 1010–1014
- 31 Tebas P, Stein D, Tang W W, et al. Gene editing of *CCR5* in autologous CD4 T cells of persons infected with HIV. *N Engl J Med*, 2014, 370: 901–910
- 32 Qasim W, Amrolia P J, Samarasinghe S, et al. First clinical application of Talen engineered universal CAR19 T cells in B-ALL. *Blood*, 2015, 126: 2046
- 33 Barzel A, Paulk N K, Shi Y, et al. Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. *Nature*, 2015, 517: 360–364
- 34 Gao F, Shen X Z, Jiang F, et al. DNA-guided genome editing using the *Natronobacterium gregoryi* Argonaute. *Nat Biotechnol*, 2016, 34: 768–773



彭智宇

北京大学生物信息学博士。华大基因首席产品官，广州市肿瘤贯穿组学研究重点实验室副主任。负责华大基因临床医学产品的研发及推广。2015年获得“广东省特支计划-青年拔尖人才”荣誉称号。主持和参与了“863”计划、“973”计划、ICGC 国际肿瘤基因组项目等重大科研项目。致力于将基因组学、转录组学、蛋白及代谢组学等多组学研究手段与临床相结合，探寻基因突变导致各种疾病的机制和相对应的个性化治疗；带领团队以大平台、大数据的方式开展出生缺陷及肿瘤的跨组学研究；参与制定高通量基因组测序的国际标准及遗传病临床表型描述的中文标准；带领团队开发基于基因组学的药物分析平台，为药物研发提供便利。

Genetic variations and human health

HUANG Hui¹, DENG JianLian^{1,2}, ZHANG XinXin^{1,2}, LI Yang¹ & PENG ZhiYu¹

¹ BGI-Shenzhen, Shenzhen 518083, China;

² BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China

In the 21st century, one of the major scientific challenges is to clarify and understand the significance of genetic variations in human health and diseases. Human genetic variations are usually correlated with the regular physiological function, so it has a tremendous boost for scientific researches and clinical testing in order to understand the effect of human genetic variations on human health. Related researches published after the Human Genome Project have enriched our knowledge of different human genetic variations and their evolutionary history and relationships. Many diseases are associated with clinical genetic variations which are prevalent in the genome. The phenotypic diversity, inherited diseases and susceptibility to complex diseases among individuals are considered to be caused by these genetic variations and non-genetic factors (e.g. environmental factors). Each genetic variation tends to have specific effects on human health. Most common genetic variations like SNPs, structural variations, chromosomal aberrations and genetic susceptibility are correlated with human health. The studies in this field will be greatly facilitated by deep understanding of the mechanisms by which mutations exert their dominant effects on human health. Moreover, understanding the effects of environmental factors and genetic variations including their interaction on human health are of the great challenging. Recent advances in technologies, especially the high-throughput sequencing technology and progressive development of related technologies, like next generation sequencing including whole genome sequencing, whole exome sequencing and targeted exome capture based sequencing have promoted scientific researches and clinical applications, which subsequently creates great improvement in the genetic diagnosis and treatment programs. Currently, the use of high-throughput technology on genetic testing has greatly expanded its scope from a single genetic disease screening to the detection of complex diseases and hereditary cancers. In present days, it is widely used in non-invasive prenatal testing, neonatal genetic testing, pedigree analysis and carrier screening. Simultaneously, technological improvement is also helping development in the field of pharmacogenomics and nutrigenomics. Genetic testing benefits research and clinical development greatly. Development and application of DNA sequencing technology has significantly promoted the study of human genetic variation and phenotypic differences relationship, and in particular to identify susceptible genes for complex diseases. Additionally, it has also contributed to studying the relationship between genetic variation and various drug resistance or sensitivity, and the specific reactions of individuals to environmental factors. This article focuses on the impact and mechanism of different genetic variations of human health, and progress on scientific research and clinical applications led by technological advances, including the prevention and treatment of genetic diseases. At the same time, we also engage with future research trend of genetic variations, hoping to provide a reference for related research on human health.

genetic variations, environment, influence, human health

doi: 10.1360/N972016-00586