

拷贝数变异的全基因组关联分析*

孙玉琳 刘 飞 赵晓航**

(中国医学科学院北京协和医学院, 肿瘤医院肿瘤研究所, 分子肿瘤学国家重点实验室, 北京 100021)

摘要 基因组拷贝数变异(copy number variations, CNVs)是指与基因组参考序列相比, 基因组中 ≥ 1 kb 的 DNA 片段插入、缺失和/或扩增, 及其互相组合衍生出的复杂变异。由于其具有分布范围广、可遗传、相对稳定和高度异质性等特点, 目前认为, CNVs 是一种新的可以作为疾病易感标志的基因组 DNA 多态性, 其变异引起的基因剂量改变可以导致表型改变。最近, 一种基于 CNVs 的新的疾病易感基因鉴定策略——CNV 全基因组关联分析开始出现, 这一策略和传统的基于单核苷酸多态性的关联分析具有互补性, 通过认识基因组结构变异可以认识复杂疾病的分子机制和遗传基础。

关键词 拷贝数变异, 全基因组关联分析, 单核苷酸多态性, 遗传标志, 复杂疾病

学科分类号 Q39, Q75

DOI: 10.3724/SP.J.1206.2008.00881

目前, 以第三代遗传标志——单核苷酸多态性(single nucleotide polymorphism, SNP)为基础的全基因组关联分析(genome-wide association studies, GWAS)已经成为研究常见复杂疾病遗传易感性的主要手段。近些年来, 利用这一方法人们已经成功地将上百个临床表型与常见序列多态联系起来, 鉴定了 200 个以上的疾病易感基因或染色体相关区段。然而, 人们随后惊讶地发现, 这些位点或区段仅仅能够解释大约 2%~15%的疾病家族聚集性原因^[1]。人类遗传学研究面临的另一个挑战是如何解释其余的遗传变异, 甚至是散发性疾病的分子基础。

早在 20 世纪 80 年代, 人们已经发现基因组中存在多种类型的染色体数目和结构变异。2004 年, Iafrate 和 Sebat 各自所在的研究小组首次在人类基因组中描述了拷贝数变异(copy number variations, CNVs)的存在^[2,3]。它是指与参考序列相比, 基因组中 ≥ 1 kb 的 DNA 片段插入、缺失和/或扩增, 及其互相组合衍生出的复杂染色体结构变异^[4]。随后, 2006 年 Redon 等^[5]在 HapMap 计划的 270 名正常健康供者中鉴定到 1 447 个 CNV 区域(CNV region, CNVR), 它们覆盖了 12%(300 Mb)的人类基因组, 而且与基因组变异和疾病致病/易感基因位点相关。这些结果提示, CNVs 可能像 SNPs 一样影响着基因的表达、表型的变异和适应, 因此也

是一种重要的疾病易感变异, 能引起疾病或增加复杂疾病的发病风险^[1,4]。因此, 在随后的两年中, 多种常见复杂疾病的全基因组 CNVs 分析结果相继出现, “CNV 全基因组关联分析(CNV association analysis)”的概念也日趋成熟^[5]。本文就 CNVs 的人群遗传学特点, CNV 全基因组关联分析的研究策略和方法及其在疾病易感基因鉴定中的应用等问题简要综述。

1 CNVs 的人群遗传学特点

了解 CNVs 的人群遗传学特点是进行 CNV 全基因组关联分析的基础。

1.1 CNVs 在基因组中的分布特点

根据基因组变异数据库^[6](database of genomic variants, DGV, <http://projects.tcag.ca/variation/>)的统计, 截至 2008 年 11 月 10 日, 基于 NCBI build 36 参考基因组序列发现的 CNVs 总共为 19 792 个, 由此形成的 CNVR 共 6 225 个, 覆盖 18.8%的人类基因组常染色质, 涉及基因 7 265 个。其中, 每条染色体上的 CNVs 数目和覆盖基因数见表 1。

* 国家高技术研究发展计划(863)(2006AA02Z19B, 2008DFA31130)和国家自然科学基金(30721001, 30772507)资助项目。

** 通讯联系人。

Tel: 010-67709015, E-mail: zhaoxh@cicams.ac.cn

收稿日期: 2008-12-25, 接受日期: 2009-05-26

Table 1 The numbers of CNVs and their involved genes in each chromosome in human

表1 人类基因组各条染色体上的 CNVs 数目及涉及基因数

染色体	CNV 数目	含 CNV 的基因数	染色体编码基因数	染色体全长(Mb)
1	1 518	724	2 107	247
2	1 402	447	1 340	243
3	1 139	344	1 096	200
4	1 411	263	777	191
5	1 000	314	893	181
6	1 073	295	1 087	171
7	1 251	399	988	159
8	1 177	288	732	146
9	985	307	848	140
10	893	257	812	135
11	921	466	1 363	134
12	898	312	1 056	132
13	509	106	353	114
14	544	170	639	106
15	774	231	655	100
16	731	396	900	89
17	668	477	1 230	79
18	426	97	286	76
19	700	615	1 453	64
20	454	193	578	62
21	255	81	253	47
22	470	224	504	50
X	524	224	868	155
Y	69	35	89	58
合计	19 792	7 265	20 907	3 079

但是, 最近的精细分析, 即用更高密度 SNP 芯片(Affymetrix SNP 6.0)对 HapMap 270 名健康个体的重新分析结果表明, CNVs 在人类基因组上覆盖的物理长度实际上远远小于 Redon 等先前在 HapMap 样本中的描述, 大片段的 CNVs(> 50 kb)只影响不到 5%的基因组序列^[6]。

尽管如此, 虽然目前发现的人类基因组中 CNVs 的个数远远低于~1 200 万的 SNPs(图 1), 但是, 它们覆盖的染色体长度至少达到 150 Mb, 这也是目前任何一种遗传标志都不能比拟的。另外, CNVs 在染色体上的分布具有非随机性, 它与其他基因组特征, 如外显子、可移动元件(如 Alu 重复序列)等密切相关, 而这些基因组特征通常是导致疾病发生的遗传学基础之一。此外, 它们的分布还具有明显的富集区和稀有区, 例如人类染色体上有 250 Mb 的区域超过 50%的序列发生 CNVs, 而有 60 Mb 的区域 90%以上的序列位于 CNV 中。这些富集区主要集中在近着丝粒和亚端粒区等具有高度多态和进化不稳定的区域^[6]。

1.2 CNV 的组成形式

目前的研究发现, CNVs 既可以是简单的

DNA 结构变化(如单一片段的扩增、缺失、插入), 也可以是复杂的染色体扩增、缺失和插入的各种组合形式(图 2)。非等位基因同源重组和非同源末端连接可能在 CNVs 的形成中发挥重要作用, 尤其是前者。Redon 等^[6]根据 CNVs 的遗传和组成形式, 将 CNVs 分为 5 类: a. 缺失, b. 扩增, c. 同一位点并发的缺失与扩增, d. 多等位基因位点, e. 复杂难以描述的位点。通常, 扩增比缺失更为常见, 并覆盖更大的范围, 这主要是因为染色体大片段缺失通常会引起更为严重的表型后果, 甚至会造成携带该变异的胎儿致死, 难以在进化中保留下来。另外一个与 CNVs 相关的概念是重复片段倍增(segmental duplication, SDs), 它是指参考基因组序列中出现 DNA 片段长度 > 1 kb 的两个或两个以上拷贝, 不同拷贝之间的序列同源性 > 90%^[4,7]。全基因组中 SDs 的密度约是 4%~5%, 而 CNVs 富集区的 SDs 平均密度约 25%, CNVs 稀有区的平均密度近 2%~3%。因此, CNVs 和 SDs 的发生具有高度相关性, 表明两者可能具有相似的发生学基础。目前认为 SDs 也是 CNVs 的一种组成形式^[7]。

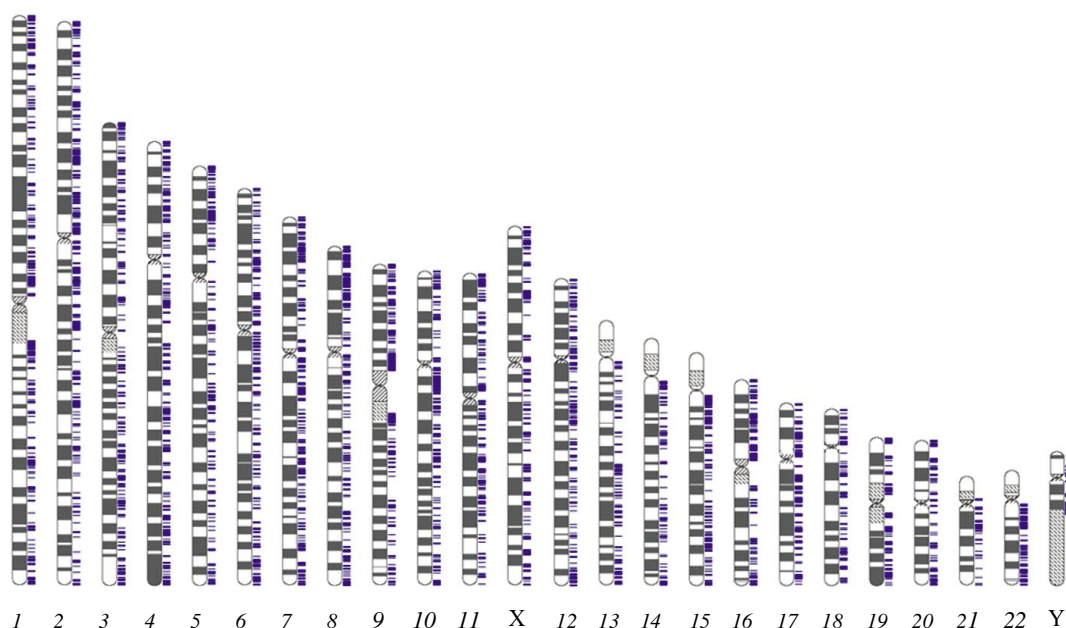


Fig. 1 The distribution of CNVs in human genome

图 1 CNVs 在人类基因组中的分布

图中每条染色体右侧的蓝线代表一个相应位置的 CNVs(数据来自 DGV 数据库).

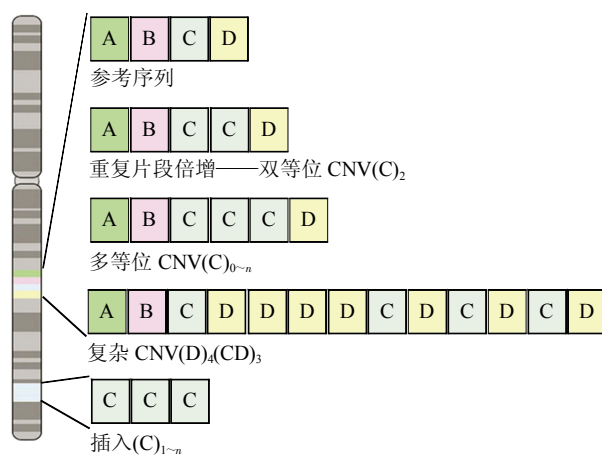


Fig. 2 Categories of CNVs^[8]

图 2 CNVs 的几种组成形式^[8]

1.3 CNVs 具有可遗传性、相对稳定性和高度异质性

CNVs 除了具有上面提到的覆盖范围广、组成形式多样的特点以外, 还具有其作为疾病易感标志的三个重要特点——可遗传性、相对稳定性和高度异质性. 近期 McCarroll 等^[6]对 HapMap 计划 270 名个体的高分辨 CNVs 研究结果显示, 正常个体中的绝大多数 CNVs 遵从孟德尔遗传规律, 而且它们

在人群之间的传递相对稳定, 符合 Hardy-Weinberg 平衡定律. 两个不同个体之间的 CNVs 变化不足 0.5%, 而只有不到 1% 的 CNVs 无法通过同一等位基因的简单遗传方式来解释. 因此, 相对于 SNPs 的概念, 将人群中等位基因频率 > 1% 的 CNVs 定义为基因组拷贝数多态(copy number polymorphisms, CNPs), 90% 以上的 CNVs 属于这一类型, 而 < 1% 的 CNVs 称为罕见 CNVs. 应用一定的算法, 可以将 CNPs 划分为双等位或多等位位点 (~10% 的 CNVs)^[6]. 其中双等位缺失位点可以有 0、1、2 三种拷贝数, 代表三种基因型, 相应的, 双等位扩增位点可以有 2、3、4 三种基因型. 而多等位位点可以有比较复杂的拷贝数变化范围, 可以通过降维 (比如将 ≥ 4 的拷贝数合并为 1 个值) 等方法来处理. 将 CNVs 理解为数量性状位点, 有利于阐述各种 CNVs 的遗传力.

另外, 利用父母-子女三人同胞对(Trios)样本分析时发现, 子女中绝大多数的 CNVs 遗传自父母, 这些位点成为遗传性 CNVs (inherited CNVs), 而新发生的与父母染色体同源序列重合率 < 50% 的 CNV, 称为新的 CNV 或新的拷贝数突变(De novo CNVs, or De novo CN mutations)^[9]. 遗传性 CNVs 通常是某些疾病具有家族聚集性的遗传学基础, 而

新的拷贝数突变可能导致某些散发性疾病的发生^[1]。

此外，目前的研究认为，CNVs 具有与 SNPs 相似程度的人种差异，即不同人群之间可能共享大部分相同的 CNVs，而部分 CNPs 在不同人群中以不同频率分离并有显著性差异^[4,6]。CNVs 具有的上述人群遗传学特点，其多态性成为一种新的遗传标记，可以用于鉴定疾病易感基因位点的关联分析。

1.4 CNVs 与 SNPs 作为全基因组关联分析遗传标志的特性比较

除了上面提到的区别以外，对于复杂疾病来讲，由于致病性变异可能分布在不同的染色体上，因此以 SNPs 为基础的关联分析对疾病易感位点的检出能力有限，即单一位点的 SNPs 等位基因无法有效地将受累个体和健康对照区分开来。然而，对于致病性 CNVs 来说，则不存在这样的问题。因为 CNVs 引起的基因剂量改变足以改变表型。所以拷贝数变异的全基因组关联分析更容易鉴定到致病突变^[10]。

目前有学者认为，SNPs、CNVs 的组成与拷贝数以及环境因子都参与疾病特定表型的产生，因此

在进行全基因组关联分析时，应当尽可能结合 SNPs 和 CNVs 基因分型结果^[10]。例如，Stranger 等报道约 18%可检测的遗传变异可以用 CNVs 来解释。另外，当通过关联分析比较 SNPs 和 CNVs 对数量性状位点的检出效率时发现，只有不到 20%的 CNVs 相关与 SNPs 相关重合，表明两种方法具有互补性^[4]。

2 CNV 全基因组关联分析的基本方法及技术路线

2.1 CNVs 全基因组扫描的方法

迅速发展的全基因组拷贝数扫描平台使 CNVs 用于全基因组关联分析成为可能。常用的技术平台有基于大插入片段的比较基因组杂交(comparative genomic hybridization, CGH)、代表性寡核苷酸微阵列分析(respectively oligonucleotide microarray analysis, ROMA)、基于长的等温寡核苷酸探针的比较基因组杂交(long isothermic oligonucleotide-based CGH arrays)和 SNP 芯片等(表 2)。

Table 2 Comparisons of common genome-wide scanning platforms of CNVs

表 2 常见 CNVs 全基因组扫描技术比较

	信噪比	分辨率 /kb	用于 GWAS 的性能
基于大插入片段的 CGH	高	> 50	低
ROMA	最高	30	中等
基于长的等温寡核苷酸探针的 CGH	中等	6~10	高
SNP 芯片	低	3	最高

2.1.1 基于大插入片段的比较基因组杂交。最有代表性的方法是细菌人工染色体(bacterial artificial chromosome, BAC)CGH 芯片，也是第一代全基因组 CGH 芯片技术。它将构建的 BAC 文库以微阵列的方式点制在片基上，用不同的荧光素分别标记正常和肿瘤组织 DNA，并与芯片上的 BAC 文库杂交，通过不同荧光素的强度比值，检测基因组中特定染色体的拷贝数变化。因为这种方法在极大程度上覆盖了基因组，费用昂贵、耗时耗力，而且因为大片段的 BACs 决定了这种方法的分辨率较低(>50 kb)^[11]。

2.1.2 ROMA。2003 年，Lucito 等建立了 ROMA 技术。这种方法的基本原理与 BAC-CGH 类似，但是它通过全基因组抽样的方法产生了一系列“代表性”的寡核苷酸片段代替了基因组 DNA，从而大

大降低了样本复杂度，增加了芯片杂交的信噪比。参考和检测 DNA 样品经限制性内切酶片段化，并在两端连上接头，经过 PCR 扩增后产生约 1.2 kb 大小的片段，代表了全基因组中约 2.5%的序列。DNA 扩增片段经荧光标记后与点制在芯片上的 70-mer 寡核苷酸探针杂交。ROMA 可以将基因组的平均分辨率提高到 30 kb。由于 DNA 样品在杂交前经过 PCR 扩增，因此 ROMA 需要的样本量明显减少。样本复杂程度降低的同时，又造成某些基因组区域的 DNA 代表性不够，而引起 CNVs 鉴定的错误。另外，不同个体可能具有不同的酶切图谱，可能造成不同个体间真实拷贝数变化的差异^[11]。此外，PCR 扩增可能存在偏好性，有时难以反映拷贝数变异的实际倍数。

2.1.3 基于长的等温寡核苷酸探针的比较基因组杂交。这类技术以 Nimblegen 公司和 Agilent 公司的全基因组 CGH 芯片为代表。芯片由具有相似热力学曲线的 45~85mer 寡核苷酸探针高密度点制而成。这些探针以迭瓦方式覆盖全部基因组序列, 包括重复序列。因此具有高精度、高灵敏度、直接检测(非“代表性”)、样本需要量小的特点, 平均分辨率达到 6 kb。

2.1.4 SNP 芯片。

SNP 芯片是目前通量最高、使用范围最广的全基因组 CNVs 分析平台。通常通过连续 3 个以上的 SNPs 探针发生缺失或扩增来定义 CNVs^[4,9,12]。由于当前的 SNP 芯片探针在基因组分布不均衡, 很多 SDs 和复杂 CNVs 区域设计 SNPs 探针较困难^[11]。因此, 最近 Affymetrix 和 Illumina 公司相继推出了 Genome-wide SNP 6.0 和 Illumina 1M 芯片, 除包括 SNPs 探针外, 还有几乎等量的拷贝数探针, 减少了因探针分布不均衡的干扰, 使全基因组平均分辨率达 3 kb, 增加了 CNVR 边界定位的精确性, 更适用于全基因组关联分析^[9]。主要优点是, 这种芯片在选择基因座上能同时提供拷贝数和基因型信息, 可以显示杂合性缺失和双亲二倍体(拷贝数中

性的杂合性缺失)等, 非常有利于寻找疾病易感基因的研究。

然而即使如此, 当前的各种 CNV 全基因组扫描技术平台仍然具有一定的局限性, 比如对于更小的 CNVs 检出效力有限(< 20 kb), 位于扩增富集区(通常是产生新突变的热点)或人类基因组中某些“新”的区域内的 CNVs 难以检测到等。

2.2 CNV 分析软件和算法

全基因组 CNVs 扫描后面面临的一个重要问题就是统计分析。目前已经发展了多种在全基因组水平推算 CNVs 的软件包和算法模型。其中比较常用的算法是隐马可夫模型(Hidden Markov Model, HMM)、环状二元分割(Circular binary segmentation, CBS)、等级分割(Segmentation algorithm)、核平滑算法(Kernel Smoothing algorithm)等(表 3)。但是, 无论哪种方法都具有一定的假阳性和假阴性率^[9]。事实上, 由于不同的样本、不同的算法、使用不同的参考样本组, 而且 CNVs 判别标准也不尽相同, 造成不同研究之间重合的 CNVs 仅介于 25%~45%之间^[13]。即使同样的研究样本, 用不同算法得到的 CNVs 重合率也只有 72%^[9]。这提示不同类型算法/软件之间可能具有一定互补性。

Table 3 The basic principles and applications of common CNVs inferring softwares/algorithms

表 3 常用的 CNVs 分析软件/算法的基本原理和适用类型

软件 / 算法名称	基本原理	分析平台类型	使用方式
Chromosome copy number analysis tool 3.0, CNAT	将受试样品组与 HapMap 参考人群或自定义的参考样品组比较, 估计每一位点的拷贝数, 并根据核平滑算法推算 CNVs	Affymetrix SNP 芯片	Affymetrix 公司, 免费
Copy Number Analyzer for GeneChip® arrays, CNAG	通过用户定义的最佳参考样品组对受试样品组归一化和测量, 并用 PCR 产物的长度和 GC 含量对信号强度比值校正, 用隐马可夫算法抽提特征并推算 CNVs	Affymetrix SNP 芯片	免费
dChip	在一组芯片中对数据进行归一化处理和测量, 并与用户自定义的参考样本组比较, 用隐马可夫算法推算 CNVs	Affymetrix SNP 芯片	免费
Gain and Loss Analysis of DNA, GLAD	此算法无法对芯片数据进行归一化处理, 但是可用于基因组断裂点的检测和 CNVs 的推算	各种 CGH 芯片	免费
Circular binary segmentation, CBS	通过比较受试和参考样本相同探针的荧光强度比值, 找到染色体上比值发生改变的位点, 而后利用环状二元分割算法将染色体分割成连续的等拷贝数的区域, 并推算 CNVs	各种 CGH 芯片	免费
Partek Genomics Suite	根据片段长度和探针序列调整粗信号强度, 同时设立参考样本组, 并用其将受试样品归一化到相同的分布水平上, 并设立拷贝数的基线值, 而后用隐马可夫模型算法推算 CNVs	Affymetrix SNP 芯片	Partek 公司, 付费
CNV Finder	适合以迭瓦方式覆盖基因组的大插入克隆的比较基因组杂交平台	大插入克隆的 CGH 芯片	免费
PennCNV	整合的隐马可夫模型算法, 将 SNP 芯片的 B 等位基因频率和 Log R 比值与探针信号强度、多种 SNP 位点人群特征和家系信息结合起来, 综合推算 CNVs	SNP 芯片	免费

续表			
软件 / 算法名称	基本原理	分析平台类型	使用方式
Quanti-SNP	用“观察的贝叶斯隐马尔可夫模型”同时将 SNP 芯片的 B 等位基因频率和 Log R 比值 (观察探针强度与期望强度比值的 log 值) 结合起来推算 CNVs	SNP 芯片	免费
Birdsuite	基于 Birdseed(SNPs 的基因分型)、Canary(CNVs 的基因分型)和 Birdseye(修正的隐马尔可夫算法推算罕见或新的 CNVs)三种算法, 提供每一位点整合的基因型信息	Affymetrix SNP 5.0 和 6.0 芯片	免费
cnvPartition	用隐马尔可夫模型同时将 SNP 芯片的 B 等位基因频率和 Log R 比值结合起来推算 CNVs	Illumina SNP 芯片	Illumina 公司, 付费
Nexus Copy Number	使用专利的等级分割算法(修正的 CBS 算法), 提高了 CBS 算法的运算速度	寡核苷酸 CGH 芯片和 SNP 芯片	BioDiscovery 公司, 付费
Helix Tree Copy Number Analysis Module, CNAM	首先用 Log R 比值结合参考样品组将受试样品信号强度归一化, 并做批间、分层调整和 Log R 值的全基因组相关分析, 利用分割算法找到染色体上荧光强度发生比值改变的位点, 从而推算 CNVs	Affymetrix SNP 芯片	Golden Helix 公司, 付费

2.3 CNV 全基因组关联分析的流程

根据文献, 目前针对无关病例 - 对照样本的 CNV 全基因组关联分析流程与 SNPs 的全基因组关联分析流程相似(图 3), 只是前者要从能够有效区分病例和对照组的 CNVR(与疾病明显相关的 CNVR)中, 按照功能候选的角度找到相应的基因, 并在独立样本中按照候选基因关联分析的方法验证分析, 从而评价候选基因对疾病风险的贡献^[12]。比较特殊的是 CNVs 特有的“新的拷贝数突变”的鉴定(图 3), 这类位点必须使用前面提到的 Trios 或父母 - 子女对样本, 通过比较子女与父母的 CNVR 差异来发现。由于当前 CNVs 研究平台和分析算法

的限制, 新得到的这类位点必须经过严格的筛选证实, 进一步去除假阳性结果。保留下来的位点通过实验验证(采用荧光原位杂交、实时定量 PCR、质谱等检测方法), 才能确定“新的拷贝数突变”, 最后还需要进一步评价这些位点与疾病的相关性以及发病风险的贡献^[9, 14]。通常这类位点非常罕见, 有的甚至可以在散发性疾病的发病中发挥重要作用。因此, 利用 CNV 全基因组关联分析, 不仅有利于寻找那些影响疾病易感性的遗传性 CNVs, 还有利于筛选人群中新的拷贝数突变, 并评价散发性疾病的遗传学基础。总而言之, 一个较为完整的 CNV 全基因组关联分析文章至少要包括 6 个部分:

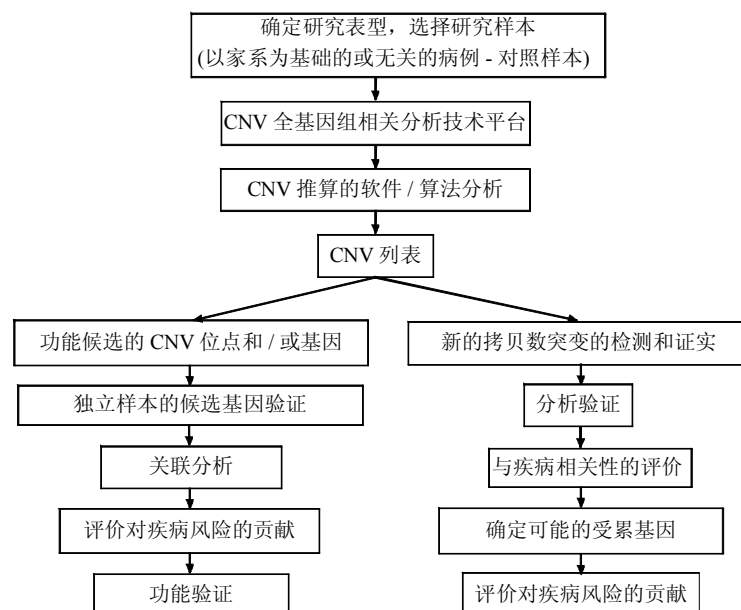


Fig. 3 The workflow of CNV genome-wide association analysis

图 3 CNV 全基因组关联分析流程

对样本的描述、CNV 检测和分析方法、实验的质量控制、CNVs 结果的描述、CNVs 与疾病表型相关性的评价以及后续的验证分析等。

3 CNV 全基因组关联分析在疾病易感基因鉴定中的应用

3.1 通过 CNV 全基因组关联分析评价新的拷贝数突变在散发性疾病中的作用

目前的 CNV 全基因组关联分析主要集中在散发性疾病与新的拷贝数突变方面(表 4)。传统认为, 遗传性疾病通常是指遗传性状与改变蛋白质结构、功能或调节的突变碱基以孟德尔遗传方式分离。然而, 临床遗传学家发现, 至少有 97% 的疾病是散

发的, 而且不涉及任何基因的突变, 仅仅源于 CNVs^[5]. 那么散发性疾病的分子基础是什么呢? 目前认为它通常是由一个染色体异常或隐性性状或新的显性突变引起的. 据估计新的基因组重组位点特异突变率介于 10^{-6} 到 10^{-4} 之间, 至少比点突变率高 2~4 个数量级. 因此对于大部分散发性疾病病例来说(甚至可能包括遗传性疾病), 新的拷贝数突变可能是一个主要的遗传性机制. 也有可能部分散发性疾病源于分布于相同或不同位点的两个不同(父母)来源的 CNVs 组合, 而仅携带其中之一的个体不会发病^[5]. 因此, 当前的 CNV 全基因组关联分析为研究散发性疾病的分子机制提供了一条有效路径.

Table 4 Published significant studies of CNV genome-wide association analysis to date

疾病类型	研究平台	CNV 定位	CNV 类型	基因	OR 值	病例组频率 /%	一般人群频率 /%	参考文献
自闭症	ROMA 和 CGH 芯片	2q24.2, 2q37.7 2q37.2-q37.3 2p16.1, 3p14, 2 6p23, 7p21.1 13q14.12-q14.13 10q11.23-q21.2 15q11-q13.33 16p13.3, 20p13 20p13-p12.3 22q13.31-q13.33	新的拷贝数 突变(缺失或 扩增)	—	—	10.2	1.0	[16]
自闭症	Affymetrix SNP 5.0	16p11.2 16p11.2	新的拷贝数 突变(缺失) 新的拷贝数 突变(扩增)	— —	100 16	1 0.56	0.01 0.04	[17]
精神分裂症	Affymetrix SNP 5.0 和 6.0	1q21.1 15q11.2 15q13.3	新的拷贝数 突变(缺失) 新的拷贝数 突变(缺失)	— — —	14.83 2.73 11.54	0.23 0.55 0.17	0.02 0.19 0.02	[14, 18]
精神分裂症	Affymetrix SNP5.0	14q32.13-14q32.2 1p34.2, 22q11.21 3q22.2, 6p22.3 16p13.2, 9p23 12q24.23, 5q31.1 3p24.1, 19q13.12 3p26.1, 4q34.1 10q21.1, 7q11.23 2q11.1-2q11.2	新的拷贝数 突变(缺失或 扩增)	—	—	9.9	1.26	[9]
动脉瘤性蛛网膜下腔出血症	Illumina 300K SNP 芯片	14q24.3-q31	遗传性缺失	SEL1L	—	21.8	10.7	[19]
骨质疏松症	Affymetrix 500K SNP 芯片	4q13.2	遗传性缺失	UGT2B17	1.73	97.9	98.9	[12]
精神分裂症	Affymetrix 250K SNP 芯片	2p16.3, 9q33.1 2p25.3, 5p15.2	罕见 CNVs (缺失或扩增)	NRXN1 ASTN2 MYT1L CTNND2	—	1.36	0.14	[20]

对自闭症和精神分裂症的全基因组关联分析就是很好的证明. 自闭症是最早开展 CNV 全基因组关联分析研究的疾病类型. Sebat 等^[16]利用 ROMA 结合 Agilent CGH 芯片分析了 118 例散发性、77 例家族性自闭症和 196 例健康对照, 结果在 16 名个体(14 名患者和 2 名对照)中鉴定到 17 个新的拷贝数突变. 这些自发性突变在散发性自闭症患者中的频率是 10.2%(12/118), 在家族性患者中的频率为 2.6%(2/77), 而在对照中的频率仅为 1%(2/196), 表明这些新的拷贝数突变是引起自闭症的主要原因. 随后, Weiss 等^[17]利用 Affymetrix SNP 5.0 芯片分析了 751 个自闭症家系和 512 名散发患儿, 发现 2 个位于 16p11.2 处新的拷贝数突变, 一个为缺失区段, 一个为扩增区段. 其中 16p11.2 微缺失在 0.35%(5/1441)家系患儿和 1%(5/512)的散发性患儿中出现, 冰岛人群中的验证分析显示, 1%(3/299)的自闭症患者具有这一微缺失, 而未受累人群的携带率只有 0.01%(2/18834). 16p11.2 微扩增在 0.49%(7/1441)家系患儿和 0.78%(4/512)的散发性患儿中检出, 而对照仅为 0.04%(2/4668).

精神分裂症是另一类受到广泛关注的复杂遗传疾病, 据报道, 它的遗传度估计在 73%~90%. 然而, 以往基于 SNPs 的全基因组关联分析并没有找到明显的易感位点^[9]. 因此, 今年有多篇文章从 CNV 全基因组关联分析的角度探讨了这一疾病的发生机制. 国际精神分裂症研究协会的多个研究小组用 Affymetrix SNP 5.0 和 6.0 芯片分析了 3 391 名精神分裂症患者和 3 181 名健康对照, 从中鉴定到 6 753 个长度超过 100 kb 的 CNVs, 健康对照平均每人具有 0.99 个 CNV, 而患者平均每人具有 1.15 个^[18]. 其中, 两个罕见的缺失型 CNVs(15q13.3、1q21.1)与精神分裂症明显相关. 随后, Stefansson 等^[19]又在 2 160 对 trios 和 5558 对父母-子女对中鉴定到 66 个新的拷贝数突变, 其中 1q21.1、15q13.3、15q11.2 三个缺失型位点与精神分裂症高度相关. 扩大样本后的验证分析显示, 这三个罕见的 CNVs 位点都明显增加了精神分裂症的发病风险, 尤其是 1q21.1 和 15q13.3. 此外, Xu 等^[20]也通过 Affymetrix SNP 5.0 芯片分析了南非白人中 359 对精神分裂症 trios, 共计 1 077 名个体的全基因组拷贝数变异情况, 在 15 名散发性精神分裂症患者和 2 名对照中鉴定到 19 个新的拷贝数突变, 它们在散发性精神分裂症患者中的携带频率为健康对照的 8 倍, 表明它们可能是散发性精神分裂症中重要

的致病因素, 然而它们并不是构成家族性精神分裂症的原因. 同时, 作者还分析了 51 个罕见遗传性拷贝数突变在散发性病例和对照中的分布频率, 两者之间的差异较小, 表明这些位点中可能只有很少的拷贝数突变是精神分裂症真正的致病因素.

上述研究证明, 自闭症和精神分裂症都是遗传异质性很强的复杂疾病, 新出现的种系基因组结构变异是引起散发性病例的重要原因. 然而, 这些结果都无法解释为什么这些疾病都具有较强的家族聚集性, 人类遗传性 CNVs 对于疾病易感性的贡献如何等问题. 也许有家族史和散发性患者之间具有不同的基因组结构变异基础.

3.2 遗传性 CNVs 的全基因组关联分析

目前, 鉴定到疾病易感性相关的遗传性 CNVs 位点的全基因组关联分析报道较少, 这主要是因为遗传性 CNVs 的鉴定比起新的拷贝数突变, 还面临着很多困难^[1, 5]. a. 现有技术鉴定到的遗传性 CNVs 通常较短, 很少超过 100 kb, 不如新的拷贝数突变那样容易被发现, 而且涉及基因较少, 结果难以解释. b. 早期的全基因组 SNPs 芯片对 CNPs 区段内的 SNPs 位点存在严重的设计偏倚, 使得这些 CNPs 很难鉴定, 新一代的 SNP 芯片有所改观. c. CNPs 以一定频率分离, 然而约 10%的多等位位点由于在人群中存在比较复杂的分离和遗传方式, 所以很难简单地使用“缺失”或“扩增”来描述. d. 新的拷贝数突变的鉴定只能基于以家系为基础的研究样本(trios 或父母-子女对), 而遗传性 CNVs 的鉴定常使用无关病例-对照样本, 因此后者可能会受人群分层等因素的干扰. e. 新的拷贝数突变往往都是患者和对照之间的明显差异, 也就是说这样的位点对于特定的疾病有较高甚至是完全的外显率, 而遗传性 CNVs 可能只有中等水平的外显率, 导致难以通过全基因组关联分析鉴定出来.

尽管如此, 目前仍然有一些遗传性 CNVs 全基因组关联分析的报道. 比如, Yang 等^[12]用 Affymetrix 500K 芯片分析了 350 名骨质疏松症患者和 350 名对照, 鉴定到 74 个 CNPs 区域, 其中 4q13.2 处的缺失型位点与疾病紧密相关. 精细定位发现其中的 UGT2B17 基因存在明显的缺失现象. 在扩大人群的重叠验证中也发现, 对照人群的 UGT2B17 基因纯合/杂合缺失率更加明显, 提示这一基因的结构变异可能是人类在进化过程中形成的保护性缺失, UGT2B17 基因的高表达增加疾病

风险 1.73 倍. Bae 等^[19]利用 Illumina 300K SNP 芯片比较了 203 名动脉瘤性蛛网膜下腔出血患者和 294 名对照的 CNVs 差异, 从中发现一个缺失型 CNV 内(14q24.3-q31)的 SNP rs1242541 与疾病明显相关, 而该 SNP 临近 SEL1L 基因. 最近, Vrijenhoek 等^[20]报道了他们用 Affymetrix 250K SNP 芯片对 54 名精神分裂症患者的全基因组 CNVs 扫描结果, 在鉴定到的 90 个 CNVs 中, 有 77 个是以前在健康人群中报道过的位点. 剩余的 13 个罕见 CNVs 中, 2p16.3、9q33.1、2p25.3 和 5p15.2(涉及 4 个候选基因)与精神分裂症相关. 在 752 名患者和 706 名对照样本中的重复验证分析显示, 这些区段在患者中的出现频率是 1.36%, 而在健康对照中的出现频率仅为 0.13%.

从上面的研究结果可以看出, 在定位遗传性 CNVs 累及的可能易感基因时, 通常有两种方式. 如果该 CNVs 本身累及了编码基因, 通常通过功能候选的方式找到可能在疾病发生发展过程中有重要作用的基因^[12, 20]. 如果 CNVs 本身位于非编码序列, 通常利用其与 SNP 位点或染色体区段的连锁不平衡来找到相应的致病基因^[9]. 遗传性 CNVs, 尤其是罕见的遗传性 CNVs 与疾病易感性的关系, 可能是未来 CNV 全基因组关联分析的新的研究热点, 将受到更加广泛的重视^[1].

4 展 望

早在 20 多年前就发现了基因组的结构变异, 近年这些结构变异与疾病的相关性越来越被人们重视. 现有的 CNV 全基因组关联分析初步展示了 CNV 在复杂疾病易感性中的重要作用. 相信随着高通量全基因组 CNVs 扫描平台和新的统计推算方法的不断发展, 基于 CNVs 这一新的遗传易感标志的 GWAS 将和基于 SNPs 及其单体型传统的 GWAS 一样, 成为研究疾病遗传易感性的有力工具. 联合使用 SNPs 和 CNVs 这两个具有互补性的遗传标志, 将为深入理解复杂疾病的分子机制和鉴定易感基因, 对遗传变异和疾病表型关系的认识具有重要意义.

参 考 文 献

- McCarroll S A. Extending genome-wide association studies to copy-number variation. *Hum Mol Genet*, 2008, **17**(R2): R135~142
- Iafate A J, Feuk L, Rivera M N, *et al.* Detection of large-scale variation in the human genome. *Nat Genet*, 2004, **36**(9): 949~951
- Sebat J, Lakshmi B, Troge J, *et al.* Large-scale copy number polymorphism in the human genome. *Science*, 2004, **305**(5683): 525~528
- Redon R, Ishikawa S, Fitch K R, *et al.* Global variation in copy number in the human genome. *Nature*, 2006, **444**(7118): 444~454
- Ionita-Laza I, Rogers A J, Lange C, *et al.* Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 2009, **93**(1): 22~26
- McCarroll S A, Kuruvilla F G, Korn J M, *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 2008, **40**(10): 1166~1174
- Feuk L, Carson A R, Scherer S W. Structural variation in the human genome. *Nat Rev Genet*, 2006, **7**(2): 85~97
- Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet*, 2007, **3**(10): 1787~1799
- Xu B, Roos J L, Levy S, *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet*, 2008, **40**(7): 880~885
- Beckmann J S, Estivill X, Antonarakis S E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet*, 2007, **8**(8): 639~646
- Carter N P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*, 2007, **39**(7 Suppl): S16~21
- Yang T L, Chen X D, Guo Y, *et al.* Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet*, 2008, **83**(6): 663~674
- Pinto D, Marshall C, Feuk L, *et al.* Copy-number variation in control population cohorts. *Hum Mol Genet*, 2007, **16**(2): R168~173
- Stefansson H, Rujescu D, Cichon S, *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature*, 2008, **455**(7210): 232~236
- Lupski J R. Genomic rearrangements and sporadic disease. *Nat Genet*, 2007, **39**(7 Suppl): S43~47
- Sebat J, Lakshmi B, Malhotra D, *et al.* Strong association of de novo copy number mutations with autism. *Science*, 2007, **316**(5823): 445~449
- Weiss L A, Shen Y, Korn J M, *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med*, 2008, **358**(7): 667~675
- The International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 2008, **455**(7210): 237~241
- Bae J S, Cheong H S, Kim J O, *et al.* Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. *Biochem Biophys Res Commun*, 2008, **373**(4): 593~596
- Vrijenhoek T, Buizer-Voskamp J E, van der Stelt I, *et al.* Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet*, 2008, **83**(4): 504~510

Genome-wide Association Analysis Based on Copy Number Variations*

SUN Yu-Lin, LIU Fei, ZHAO Xiao-Hang**

(State Key Laboratory of Molecular Oncology, Cancer Institute & Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100021, China)

Abstract Copy number variations (CNVs) refer as a DNA segment that is 1 kb or larger and is presented at a variable copy number in comparison with a reference genome. Classes of CNVs include insertions, deletions, duplications and their complex combinations. Because they widely distributed in the genome with some important characteristics, such as heritable, relative stable and heterogeneity, CNVs are considered as novel genomic polymorphism markers. And the alteration of gene dosage which resulted from CNVs could change phenotype, so a novel CNV genome-wide association analysis (CNV-GWAS) strategy appeared recently and began to used for identifying susceptible genes of complex diseases. It was approved that it could complement the traditional genome-wide association studies based on single nucleotide polymorphisms. Therefore, genomic structure variances are favorable for revealing the molecular mechanisms and genetic foundation of complex diseases.

Key words copy number variances, genome-wide association studies, single nucleotide polymorphism, genetic markers, complex diseases

DOI: 10.3724/SP.J.1206.2008.00881

*This work was supported by grants from Hi-Tech Research and Development Program of China (2006AA02Z19B, 2008DFA31130) and The National Natural Science Foundation of China (30721001, 30772507).

**Corresponding author.

Tel: 86-10-67709015, E-mail: zhaoxh@cicams.ac.cn

Received: December 25, 2008 Accepted: May 26, 2009