A copy number variation map of the human genome

——人类基因组的拷贝数变异图谱

# 研究背景

近些年的研究，发现人类基因中存在大量的中等片段长度(intermediate size)的结构变异，在DNA含量缺失或增加的突变中占很大的比重，并将其称为拷贝数变异(copy number variation, CNV)。拷贝数变异是指染色体显带不能识别的大小在50bp到数Mb的基因组之间的DNA片段的不平衡重排。

CNV的表型从可适应的特征(adaptive trait)到胚胎致死，效应范围广泛。CNV表型的划分和临床医生对相关疾病的定义有关，而且同一个CNV在不同环境条件可能是良性（适应性良好）也可能是致病性（适应性不良）。目前，适应不良的CNV与孤独症、精神分裂症、克罗恩病、类风湿性关节炎、糖尿病I型和肥胖症等多种疾病有关，并且相关CNV逐渐成为对应疾病的诊断依据。而这些因基因组结构特征导致基因组重排所致的基因组拷贝数变异引发的一类疾病称为基因组病（由一系列特殊的CNV在其重排断点有不连续的重复DNA片段延伸引发的疾病称为基因组病）。临床遗传学家的一项重要工作就是通过健康人群的CNV数据作为参考，从病人的众多良性CNV中辨别出致病性或高概率致病性的CNV。而当前该领域的主要挑战是从不断增加的未知临床特征的CNV中发现相关的致病CNV。

经过多年的发展和相关数据库（本研究主要使用DGV数据库，DGV收录健康人群的CNV和结构变异信息。DGV含有健康对照人群相对于特定病人的信息、健康组织中体细胞突变的CNV信息、人群中CNV的发病年龄等，当前数据库面临对复杂CNV特征描述和正常人群与病人之间临床表征区分等挑战）的逐渐完善，该研究致力于使用现有的信息构建一幅正常人群的人类基因组的拷贝数变异图谱(CNV map)。

本课题的主要工作就是通过对公开可获取的高质量数据进行CNV的荟萃分析(meta-analysis)得到相应的拷贝数变异图谱，首次进行基因组图谱构建的目的是为临床医生或科研人员进行相关的诊断分析和研究提供参考依据。需要指出的是，该图谱只包含相关的缺失和重复突变等不平衡结构变异，不包括倒位易位等平衡结构变异，也不包含线粒体基因突变。本图谱构建方法是分别对基因组缺失突变(deletion, loss)和基因组重复突变(duplication, gain)进行各自的构建，然后将两个子图谱融合形成统一的图谱。在构建过程中，研究人员使用到了各种方法和研究数据会在后文详细介绍，并且会呈现图谱的结构、性质和拷贝数变异区域(copy number variable region, CNVR)对基因及其功能的影响。

# CNV检测技术

微阵列和二代测序是目前主要的CNV检测方法。早期的研究主要使用BAC(bacterial artificial chromosome-based)微阵列芯片技术和寡核苷酸阵列，随后的微阵列芯片有CGH(comparative genomic hybridization)和SNP-based阵列。随着二代测序技术以及从相关数据中检测CNV信息的工具出现，近些年来该方法得到全面推广。

目前，还没有一个检测策略能够捕捉基因组中全部各种大小的结构变异，CNV探查仍然十分依赖各种使用的平台和算法。多数检测平台由于缺乏足够的片段重复序列(segmental duplication)探针，所以更加偏向于发现结构变异(structural variation)，进而很难识别许多的致病性CNV。由于测序平台不同，所得到的突变长度和分布也不尽相同。基于微阵列的检测方法更加适合于定量突变体(quantitative variants)的研究，而早期的方法(如BAC和CGH)由于低分辨率和断点过估计的特点对CNV含量会预估过高，并且错过很多小突变。另外，缺失通常比较容易检测，缺失和重复的检测方法有所不同（比如使用SNP-based阵列和二代测序技术相较于CGH技术会丢失更多的重复片段），千人基因组计划的第一阶段数据就很有可能丢失了私有的或多发但罕见的突变。

突变体的大小因平台而异。相较于微阵列技术，基于NGS测序技术的研究可得到更短的突变以及更倾向于探测到缺失突变。基于测序技术的方法灵敏度更高且能提供准确的断点位置，相较而言微阵列技术就受制于低分辨率（高分辨率微阵列检测的CNV最小下限为450bp）。重复突变的检测更加倾向于使用CGH技术，而非测序方法，因为该方法对拷贝数微小差异(small differences)更加敏感。

早期的测序研究致力于同时对单人个体即对特定基因组区域加以测序又对全基因组开展测序，而千人基因组计划是首个对来自不同人群的多样本进行全基因组测序。过去数十年CNV数据的积累为本次全面的荟萃分析提供了条件。

# 现存CNV研究成果的评估

## 研究的课题筛选

DGV数据库(Database of Genomic Variants)中收集并整理的拷贝数变异以及拷贝数变异区域来自于55个课题研究(study)。取所有这些整理好的拷贝数变异，首先根据相关课题使用的检测技术分成1)测序技术所得数据2)微阵列技术所得数据（寡核苷酸阵列或SNP-based阵列）3)其他方法所得数据（如荧光原位杂交技术FISH、PCR、多重连接探针扩增技术MLPA和光学标测技术等）等三大类。

全基因组分析和精确的断点分辨率是成为本次研究数据的主要筛选依据。1)如果微阵列平台至少有一百万个探针由此产生的数据将入选,而低分辨率的微阵列数据将抛弃。2)如果微阵列为定制的或特定目标区域的阵列，即使不是全基因组的数据，只要数据是高准确性和高断点分辨率的也将采用。3)如果某课题使用一种以上的技术得到相关数据，本研究将采用标准最低的方法所产生的数据。4)不符合上述条件的低灵敏度、低分辨率、非全基因组测序并且断点不明确的数据均将被抛弃。

经过一系列的筛选，最终55个课题研究保留下26个，其中主要的数据是通过NGS测序平台获得，然后是微阵列技术得到，还有三个是利用Sanger测序作为主要检测方法。

## 单个课题的评估

进一步对筛选后的数据加以评估，尤其注意假阳性突变和声称为罕见突变的数据。对26个课题数据各自进行突变频率统计，并且每组数据构建（至少）两个数据子集。第一个子集包含所有突变，第二子集为过滤掉低频率突变后剩余突变。低频率突变的认证标准为该组数据中含有超过100个样本数，突变频率<1%的认定为低频率突变，需剔除；该组数据中含有超过20个样本数，突变频率<5%的认定为低频率突变，需剔除；该组数据中样本数不超过20的不需要进行过滤，全部采用。这样每组数据就产生了三大类的数据集（分别是全部突变、剔除1%频率突变和剔除5%频率突变）。对各自数据集中的重复突变进行融合merge（使用最小极限50%的重叠互补簇算法;minimum 50% reciprocal overlap clustering algorithm），形成共有突变(consensus variants)，检测相关的灵敏度（两个金标准的突变数据集:Conrad validated variants和Complete Genomics variants，以及CytoSickkids）和准确度（是否和蛋白编码序列重叠，尤其是剂量敏感型基因或和疾病相关基因。其中，对各组数据进行基因类型(Decipher\_gain/Decipher\_loss、ISCA\_gain/ISCA\_loss、Essential/No-phen、CGD\_dom/CGD\_rec、 HPO\_dom/HPO\_rec、Cancer\_gain/Cancer\_loss、OMIM、…)分类和基因组元件(CpG、segmental duplication、enhancer element、putative promotor、ultra-conserved element、phastCons element、LincDNA…)分类，这些特殊类型的基因被认为是对CNV有负选择(negative selection)。）

观察图表可知，某个课题数据中所有突变的数据集与DECIPHER数据库有14.6%重叠；经频率1%过滤后，重叠率下降2.5%；再经频率5%过滤后重叠率，变化不明显。同样的现象也在与ISCA和OMIM数据库比较中能发现。

经过一系列质控，最终26个课题数据有23个用于拷贝数变异图谱构建。被排除中的三组数据中：1)Xu2011数据在与编码蛋白基因和突变不容忍(intolerance)基因的重叠率最高，并且全基因组的突变率也最高，但是经过频率1%的过滤之后重叠率显著下降，被严重怀疑可能含有大量假阳性突变，故舍去。2)考虑到低分辨率和低验证率(validation rate)，有充分的证明支持去除Itsara2009数据。3)和Xu2011数据类似，Banerjee2011有着最高的剂量敏感性突变率，并且该组数据采用了其他课题研究中没有使用的SNP阵列技术，进而很有可能引入假阳性突变，故也被剔除出去。另外，对千人全基因组计划第一阶段(1KG-Phase1)的数据加以频率1%的过滤之后，基因组突变率也出现显著下降。但是考虑到该组数据采集自15种不同人群的多样本，可能可以从中获得罕见的或特定人群的突变，所以实验人员保留了该计划的所有突变寄希望能发现不同人群之间的所有等位基因突变型。

总结可得，来自23个课题研究数据的2647个多种族样本的2057368个CNV片段将用于后续拷贝数变异图谱构建（详情见补充材料）。

# CNV图谱的构建

CNV图谱的目标是记录来自不同人群健康个体的人类基因组中的拷贝数变异。为了最大限度地构建人类CNV图谱，研究中将不同的课题研究的数据整合到单一的图谱中。相关CNV 可分成常见CNV(common)和单一CNV(singleton)。常见CNV可在不同的人群和种族中被检测到；而单一CNV指该CNV是来自单一课题单一受试者，有可能是稀少、特有或者根本就是错误，所以需要对单一CNV进行专门的解释。另外，不同平台对突变体的起始和终止位点的确认准确度不同，所以通过不同方法得到的同一突变体的坐标可能不一致。

将不同课题的高可信度CNV融合到各自的拷贝数变异区域(CNVR)，本研究采用如下的策略：

首先，由于来自不同课题的CNV可能有不同的CNV边界，故采用CNV拷贝数变异区域形成簇算法(CNVR clustering algorithm)，将互补重叠率至少为50%的片段划为一簇。

对簇的形成和选择也有两种不同的策略：一是（样本-课题法subject-study approach）不使用之前基于突变频率的过滤过程，根据重叠率直接形成簇，然后考虑每个簇内突变体的受试者和课题数，对已形成的簇加以选择，这将对不同的共有突变有不同水平的支持(consensus variants with different levels of support)；一是（打分算法scoring algorithm）先将频率<5%和1%的突变体过滤，然后依据重叠率形成簇，并根据课题中的受试者数量对每个突变体进行加权打分，接着根据簇中所有突变体的分数进行筛选。

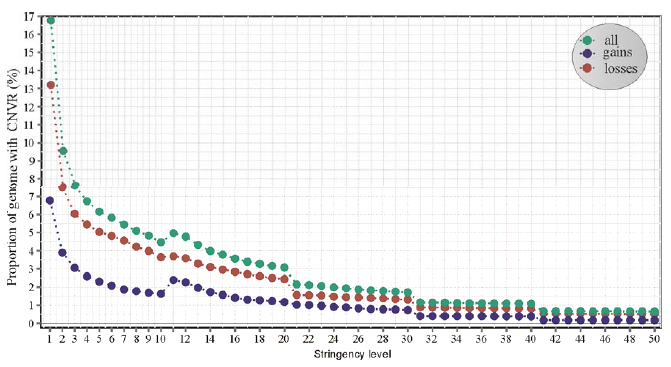
当然无论使用上述哪种方法，后期都要对各簇内的突变体加以融合，形成统一的CNVRs。融合后，还可以通过计算某CNVR的边界平均值来评估边界是否延伸过长，不过本研究所用数据并没有出现这种情况。

## 受试者-课题法(subject-study approach)

该方法需要设置受试者数和课题数两个过滤参数，并通过过滤得到最终的簇。基于受试者数的过滤可以将单一CNV排除在外，突变体由大量不同的样本数据支持将尽可能地降低假阳性突变，基于课题数的过滤可以将来自不同课题实验的人为偏差排除在外。

研究人员评估了50对不同的受试者数和课题数过滤参数的过滤结果。从总体趋势可看出，参数设定越高，得到的CNVR可信度越高；反之，如果设定每个突变体只需要来自1受试者1课题，即严密水平1级(stringency 1)，包括单一CNV在内的所有CNV和CNVR都将被采用。

从图表可知（列举其中一个如下），当严密水平从1级增加至2级（2样本1课题）时，突变率均出现显著下降（原因可能是排除了单一CNV）。当严密水平增至11级（1样本2课题）时，突变率较严密水平10级（10样本1课题）有明显增加，当严密水平增加至12级（2样本2课题）时，突变率又有所回落（原因可能是：当严密水平11级时，由于只需要1样本可能存在假阳性情况）。



## 打分算法(scoring algorithm)

**本部分在正文中不存在，详见需查看补充材料。**

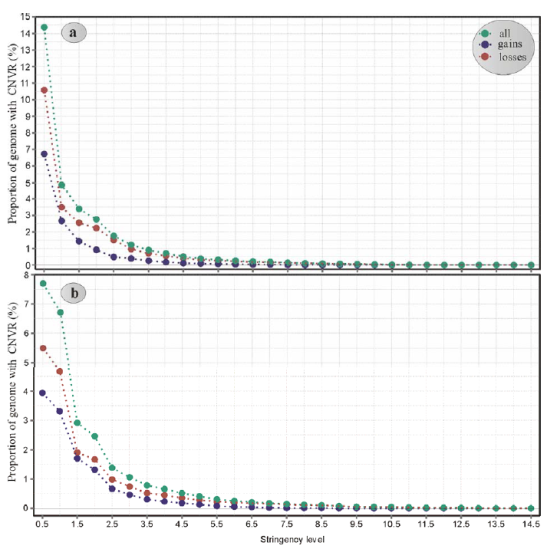
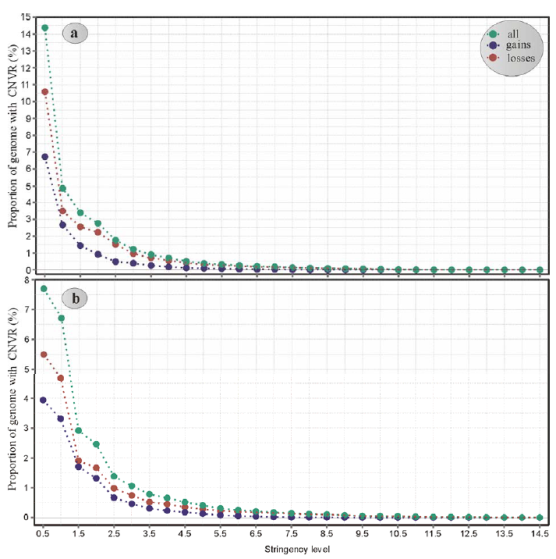
打分算法是给每个突变体打分，每个簇的分数是将该簇内的所用突变体的分数相加所得，而每个突变体的分数是依据经突变频率过滤后课题中剩余受试者数决定的。

经频率<5%过滤后：1-20受试者数的课题为0.5分，21-270受试者数的课题为1分，271以上受试者数的课题为2分；经频率<1%过滤后：1-100受试者数的课题为0.5分，101-270受试者数的课题为1分，271以上受试者数的课题为2分。和前面的课题筛选时相同，如果课题中受试者数不足以进行频率过滤，则跳过该步骤直接作为已过滤课题进入打分步骤。

依据簇的分数进行簇的筛选和图谱构建的方法是，追踪每个簇内最低突变体分数。具体来说，如果该图谱的严密水平为0.5级，则说明图谱中包含的簇的最低突变体分数为0.5分；图谱的严密水平1.5级，则说明图谱中包含的簇的最低突变体分数为1.5分。研究人员主要采用严密水平0.5级和1级进行分析。

显然，通过打分算法方法所得到的严密水平0.5级的图谱所含突变数比通过受试者-课题方法所得到的严密水平1级的图谱所含突变数要少，原因是前者有相应的突变过滤，而后者则包含所有课题中的所有突变。

从图表可知（列举其中一个如下），频率5%过滤（b图）所得的全基因组变异率普遍小于频率1%过滤（b图）的结果。突变率最为显著的下降出现在频率1%过滤后严密水平从0.5级到1级的阶段。频率5%过滤的数据出现突变率显著下降的情况发生在严密水平从1级到1.5级的阶段，类似情况不仅发生在全基因组突变率图表中，也发生在其他类型（Conrad2010 validated variants、Complete Genomics variants 、disease genes、evolutionary constrained genes）突变率图表中。两种过滤条件得到的图表出现不同的表现，很可能是因为频率5%过滤条件在去除低频率突变时没有1%的严格。



## 最终选择

为了后续研究，研究人员考察了各种条件和严密水平下构建的图谱，基于对基因组拷贝数变异的评估、来自不同平台数据的考虑和参考突变数据的重叠情况以及高度保守的基因组元件较少突变等多种因素，选择了两种过滤参数：1)每个突变体至少来自2受试者1课题，即严密水平2级，形成一幅标准相对宽松的包容图谱(inclusive map)，2)每个突变体至少来自2受试者2课题，即严密水平12级，形成一幅标准相当严格的严格图谱(stringent map)。

由簇中序列坐标均值**(using the average of the coordinates of the clusters)?**计算可知，在包容图谱中CNV占比约为8.8%（其中缺失6.9%，重复3.5%），在严格图谱中CNV占比约为4.1%（其中缺失3.1%，重复1.9%）。对各平台的断点分辨率进行排名分析，总基因组突变率与断点分布相似。**(We further ranked the platforms based on the breakpoint resolution with the Sanger and NGS platforms received the highest rank. The total genomic variability was similar to the maps where breakpoints were calculated by averaging over all variants.)?**

早期的研究证实在DGV数据库中的单一CNV不是极为罕见，就是错误发现，进而应该将其从CNV图谱中剔除。通过剔除这些单一CNV，研究人员充分相信能很好地降低数据的假阳性率和潜在的罕见致病突变体，这些罕见突变体大部分都发病较晚而且是在特定课题研究中才被发现。

# CNV图谱的性质

包容图谱有3132个重复CNVR和23438个缺失CNVR，9.5%的基因组和缺失(7.5%)或重复(3.9%)有关。早期研究预估人类基因组中全部的CNV占约为12%-13%，而大片段（长度＞50kb）的CNV约占少于5%。基于Conrad2010 validated variants数据研究可得到基因组的3.7%为CNV，该比例略低于本研究中严格图谱的比例(4.8%)，原因可能是Conrad2010数据只采集了欧洲和非洲人群样本，而包容图谱比例很高的原因是本研究的数据来自非常广泛的人群。另外，当前采用的荟萃分析所用数据主要来自于测序技术，该方法能够比传统的阵列技术捕捉到更小片段的突变。

严格图谱中的突变更加的普遍，即受种群约束的条件更小。对于突变体的严格筛选要求可有效降低平台偏差(bias)，同时也去除了很多有效且稀有的突变体。严格图谱的CNV比例为4.8%（其中缺失3.6%，重复2.3%），低于包容图谱对应比例的一半。

另外，CNV数量和突变区域的长度呈负相关变化。大部分的CNVR长度为300-3000bp，且缺失片段的长度略微小于重复区域的。

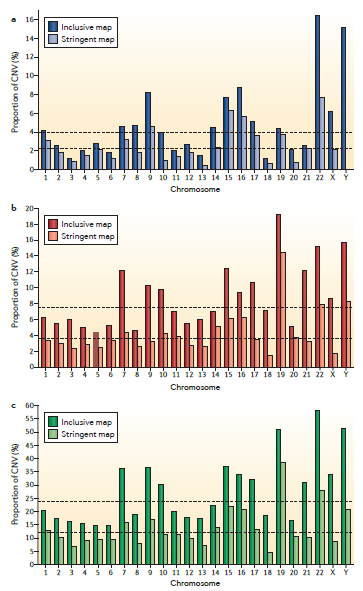
CNVR在各染色体之间分布不均匀（具体结果见下图）。

**CNV重复：**包容图谱中，各染色体上CNV重复比例从1.1%至16.4%不等。Chr22和ChrY中含有最高的突变比例，紧随之后的是Chr16、Chr9和Chr15，以上染色体中的突变比例均大于平均值；而Chr3和Chr18在该图谱中的突变比例最小。

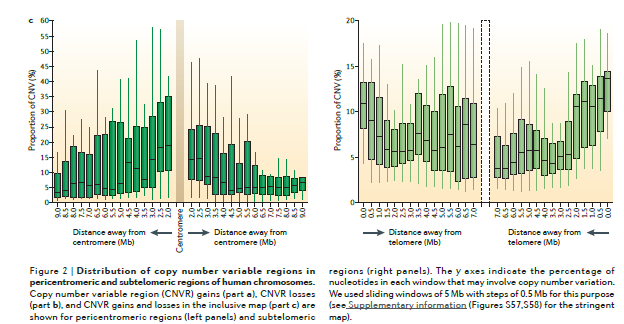
严格图谱中的重复情况和包容图谱的类似，但是ChrY表现为拷贝数稳定(copy number stable,CNS，指该区域基本不存在CNV片段)，原因可能是很难针对ChrY中高度重复区域设计出合适的测序引物，并且很多市面上出售的阵列芯片也不检测这个区域。

**CNV缺失：**包容图谱中，各染色体上CNV缺失比例从4.3%至19.2%不等。Chr19、Chr22和ChrY在包容图谱和严格图谱中均含有最高的突变比例，而Chr5和Chr8在包容图谱中的突变比例最小，Chr18在严格图谱中的突变比例最小。

**CNV缺失+重复：**将两种类型的CNV图谱合并，其分布状况和CNV缺失图谱的分布状况相似。



CNV在各自染色体上的分布也不均匀，其中着丝粒周缘区域和亚端粒区域的分布比例较高。即便由于这些区域的复杂性而导致缺乏合适的测序方法和阵列芯片，进而影响了这些区域的测序结果不如意，但是着丝粒周缘区域和亚端粒区域仍表现为较高分布情况（实例图如下）。



# CNV功能区的影响

不同的基因组元件对CNV的约束力有所不同。本研究中主要分析基因外显子序列的突变，同时还分析了转录本、编码DNA序列(CDS)和内含子的突变情况。研究中，基因根据如下条件分成不同的集合：1)编码蛋白或非编码蛋白，2)在人类、小鼠或孟德尔遗传病中有已知表型，3)基因组保守和进化约束指数(conservation and evolutionary constraint indices)；同时也对基因进行如下方面的评估：1)保守区，2)基因表达调控区（核心启动子、增强子、CpG岛），3)结构判断（着丝粒、端粒、片段重复片段和重复区域）。

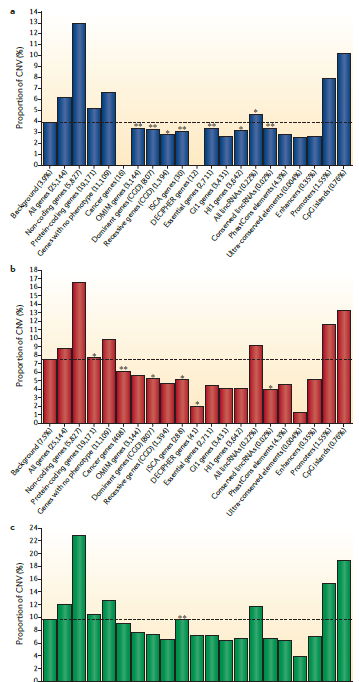
**本段落来自补充材料：为了检验结果的统计显著性，研究人员随机生成一万个样本的数据集，对基因数据和调控元件数据进行基于单向置换测试的经验P检验和二项分布P检验。结果显示，大多数基因的内含子片段比外显子片段更加稳定，也就是说内含子片段有更高的负向选择，不过该结论有待后续研究。**

从突变的总体来看，拷贝数缺失比拷贝数重复在不同基因中占比更大，原因之前已提及：CNV检测技术上存在根本性的偏差(bias)。外显子（RefSeq定义）的突变情况高于基因组的平均水平。其中外显子中的非编码蛋白基因的突变率最高（高于编码蛋白基因的外显子突变率）；与之对应的是，保守区域（尤其是和疾病OMIM,ISCA and cancer 或者其他可变适应表型fitness-altering phenotype有关）的外显子突变情况低于平均水平。

癌症相关基因和DECIPHER关键基因存在强负选择，后者存在最少的缺失突变。在DECIPHER关键基因之后，缺失和重复突变都最少的基因是由GI(genic intolerance score)和HI(haploinsufficiency index)预测而来的。在包容图谱和严格图谱中，OMIM基因的外显子显示和CNVR的重叠较少，他们的约束性和ISCA基因类似；必需基因的外显子突变和OMIM基因的情况相当。显性基因的外显子突变情况并没有显著得高于隐性基因的，早期有研究表明显性和隐性基因的CNVR相对较少（本研究证明了这一点）。

另外，lincRNAs富含CNVR，启动子相对于全基因组而言CNVR含量也较高，这种现象可能说明在基因收尾两端富含片段插入或缺失的现象。增强子在CNVR中含量很少(impoverished)，PhastCons元件和超级保守子(ultra-conserved elements)在CNVR中含量由于强负选择也相对较少。CpG岛和基因组的平均水平相比突变率较高。

过去的研究表明CNV和基因密度之间呈正相关，而本研究发现情况并不通用。片段缺失由于特定的基因负选择而含量明显较少，然而，片段重复致病概率更小，往往在正选择下驱动多组基因家族演化进程(evolution)。尤其是，影响编码蛋白基因的常见缺失(common deletions)比罕见突变还少。



# CNV和片段重复序列

片段重复序列，又叫低拷贝重复序列(low-copy repeat)。DNA中长度超1kb的高度同源重复序列之间表现出>90%的序列相似性。

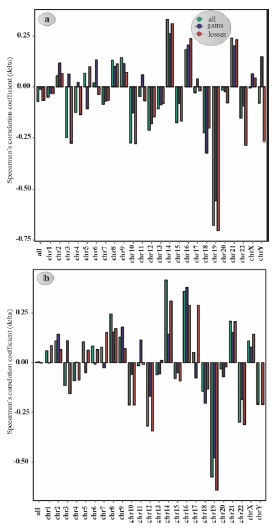
片段重复序列和基因组稳定性及疾病的关系在过去的其他综述中已经有所阐述，此类序列主要在着丝粒周缘区域和亚端粒区域富集。不同染色体在片段重复序列的含量上各异，Chr19主要含有演化更加久远的片段重复序列（较低的序列一致性，约为90%），而Chr14主要含有演化时间较短的片段重复序列（很高的序列相似度，约为99%）。

在包容图谱和严格图谱中，核酸序列中片段重复序列分别占60%和40%。CNV和片段重复序列之间的实质性重叠和关联性表明这些区域有着动态变化的性质，这将模糊了CNV和祖先性重复片段（ancestral duplications，这些片段被认为是特定人群固有的）之间的区别。低序列一致性的片段重复序列和全基因组的相关性比高序列相似性的要高，这很有可能是因为测序技术偏差(bias)所致，而非内在某种机制决定的。

**以下段落来自补充材料：研究人员采用了滑动窗打分的方法分析SD(segmental duplications)和CNVR之间的相关性。设置大小为2.5Mb的滑动窗，以0.25Mb为单位向前移动，计算每次窗口内SD的长度和一致性，后者对应于UCSC SD表中定义的fracMatch分数除以窗口长度(2.5Mb)，另外还需要计算窗口中的核酸突变率，即突变分数。在统计过程中，研究人员去除了着丝粒和异染色体区域，以及重叠区域超过50%带有装配间隙(with assembly gaps)？的窗口。结果显示在染色体上突变分数和SD分数不均匀，并且两者的增加之间呈现正相关。**

**研究人员还分析了低序列一致性和高序列相似性和CNV之间的相关性，片段重复序列选择长度为64Mb的所有高序列相似性(99%-100%)和低序列一致性(90%-94%)序列。结果显示：高序列相似性的序列和包容图谱的正相关系数分别是****0.37（重复）和0.30（缺失），和严格图谱的正相关系数分别是0.35（重复）和0.29（缺失），P值均小于0.05；低序列一致性的序列也有类似的情况（包容图谱：0.39重复、0.37缺失，严格图谱：0.34重复、0.29缺失）。**

**研究人员还通过将每个染色体上的高序列相似性系数值减去低序列一致性系数值，得到相关系数的增量分值(delta score)。通过增量分值可看出，部分染色体在片段重复序列不同类型的含量上存在显著偏差。结果如前所述：Chr19主要含有演化更加久远的片段重复序列（较低的序列一致性，约为90%，增量分值约为-0.7），而Chr14主要含有演化时间较短的片段重复序列（很高的序列相似度，约为99%，增量分值约为+0.3）。同样如前所述，由于相似度过高的片段难以设计引物导致存在测序技术偏差，所以增量分值为正数的结果普遍偏低。具体结果如下：**

****

# Null CNV图谱的构建

研究人员通过汇集在DGV数据库中被鉴定为纯合缺失的所有CNV（即为拷贝数为零），并将他们框定在相应区域，统计相关区域在包容图谱和严格图谱中正好呈现缺失型CNVR，产生人类基因组的无效(null)CNV图谱。

在严格图谱中，这些CNVR区域大小适中但不显著，相关区域富含具有旁系同源的基因，无效CNV在严格图谱中占基因组的0.75%。在包容图谱中，研究人员发现了107个人类蛋白编码基因（其中99个同时存在于严格图谱中），这些基因85%以上的外显子在无效CNVR区域内存在缺失。

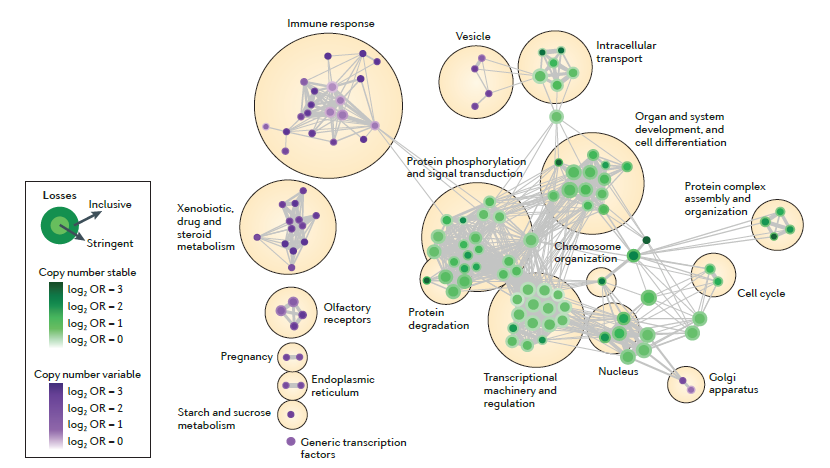
由于能够从相当健康个体的基因组中消失，这些基因有可能是非必需基因。据分析，这些基因有着很多的基因家族成员，基因的复制常见，片段重复比其他基因更加普遍。这些基因的功能可能是多余的，或者与一些迟发性疾病有关，而这些症状不会实质性地影响人体的健康水平。在这些非必需基因中，包含OMIM中和年龄相关的致病基因，如UGT2B17（骨质疏松症）、Rh阴性血型RHD、延迟HIV1感染后艾滋发病KIR3DL1、疑似多发性硬化症的基因HLA-DRB1、自身免疫疾病HLA-DQA1和 胎盘功能胎儿健康PSG1。有意思的是，本列表中的基因和MACArthur et al.报道的功能缺失基因(loss-of-function genes)没有重叠。

# 功能富集分析

研究人员通过GO和其他通路相关的基因数据库分析了易突变和稳定基因的富集情况。在包容图谱和严格图谱中，这些富集基因数据集在片段缺失和重复上情况十分类似；严格图谱表现出显著的富集情况，尤其是在稳定区域的功能性基因集中。基因的富集情况可以通过网络结构的形式可视化，以便高度重叠的基因数据集组成相关的功能簇。

在包容图谱和严格图谱中，CNVR富集的功能区有异常生物体、药物和类固醇代谢，免疫应答，嗅觉受体，淀粉和着糖代谢，妊娠特异性粘附分子，内质网，囊泡和高尔基体；CNS（拷贝数稳定）富集的功能有蛋白质磷酸化，信号转导，蛋白降解，转录机制和调节，细胞内转运，器官和系统发育，细胞分化和细胞核区室。具体情况可参见补充材料。

**基因家族中的成员间和片段重复序列重叠，这个基因家族富含CNV。(Gene families with members that overlap with segmental duplications were enriched for CNVs?)**举例来说，人类编码催化淀粉消化的酶的唾液淀粉酶基因表现为广泛的CNV。



# 讨论及后期工作

CNV图谱显示并量化人类基因组中CNV的异质性程度。人类的基因组通过结构上的不断修饰而发生着变化，这些改变将反映在健康或患病个体中。存在于功能序列之间或之中的CNVR将给良性和致病性CNV的区分带来挑战。包容图谱中，基因组中CNS区域可能是全基因组中最剂量敏感性的区域，该区域内的CNV很有可能与疾病或者适应性低下(reduced fitness)有关。在包容图谱和严格图谱中，CNV和疾病相关基因有所重叠，可以根据相关疾病的临床表征反映该CNV突变减少的外显情况。

该图谱的价值体现在将来自不同课题研究的数据汇总并加以分析，克服测序技术和微阵列芯片技术各自的局限性。另外，本课题还克服了先前研究中样本数量和采样人群不足的局限性。相比之下，千人计划分析了广泛人群中大量的样本，但是主要采用低覆盖率的测序技术，导致发现的CNV片段主要是长度小于400bp的拷贝数缺失，而拷贝数重复和大片段的CNV就显得数量不足。虽然通过荟萃分析已产生迄今为止最为完备的CNV图谱，但是长度为50bp-500bp的CNV仍需要通过高精度测序技术加以进一步的特征注释。该CNV图谱可以作为表型突变的遗传分析的比较标准。并且它将会成为CNV注释以及CNV与种群演化适应性关系鉴定的工具。另外，该图谱还能用于指导特定CNV基因型测试芯片的开发。

**临床应用：**该图谱能够作为医学中相关CNV研究的工具。在评估某病例中发现的CNV临床表征时，其标准应该包括：1)是否在该CNV图谱的CNVR中发现，2)是否和临床疾病相关性基因有重叠。研究中，人为审核并总结了935种与临床相关的基因，该全面的疾病相关基因表格来自407篇公开发表经同行审议的文献，以及ISCA剂量敏感性基因 、Baylor College of Medicine 105k 寡核苷酸芯片疾病信息、DECIPHER关键基因和癌症基因数据库。如果某病人中的CNV与上述材料中非基因有重叠，却不存在CNV图谱中，那么该CNV将有很重要的临床意义；特别的是，该CNV和DECIPHER关键基因或胚胎致死基因等严格的基因集相关，则有很高的致病概率。包容图谱和严格图谱间的差异可以考虑相关突变的条件性表型，如年龄相关表型。相较于早发性的基因，具有迟发性致病突变的基因的保守性更小，并且有更大可能性从严格图谱中过滤掉。

目前，在我们认为是正常性状和疾病之间，以及被分类为良性（或中性）的突变与被分类为易感风险因素（或疾病相关的突变）之间尚未有明确的界限，包括种族在内的多种基因组背景将影响相关界限的阈值。比如，低拷贝数的淀粉酶基因对亚洲人群是不利的，尤其是日本人，但是对非洲人群却没有明显效应。

**未来研究方向：**CNV图谱的构建会受到一些因素的制约。

首先，来自健康个体的CNV信息不够完善。当前基因组CNVR的包容图谱构建是由2647个受试者至少两个相互独立的不同样本产生的数据而来。随着更多可用的CNV数据生成，图谱将会进一步完善。但是，欧洲人群的CNV数据接近于饱和，因为多增加近三千该人群的数据对基因组突变体的数量增加影响微小（从9.5%增加至9.7%）。

很多在图谱中被认定为是CNS区域，其中很多区域可能真得含有突变，因为低频率的突变只有在较高人群基数下才能显露出来，或者CNV在不同人群中的频率不尽相同（在受试者人群中该CNV相对稳定）。举例来说，在片段重复序列区域中的CNV比特有序列区域中的CNV更可能有群体分层性。1)Campbell et al.的研究表明非非洲人群的苦味味觉受体基因(TAS2R46、TAS2R48)中含有CNV比非洲人群更加普遍，2)本研究还发现与闭合蛋白基因(OCLN)有关的CNV在东亚人群中的拷贝数要低于非洲人群。3)趋化因子配体3样1基因(CCL3L1)是拥有外显子CNV（不同人群中）区分最明显的基因之一。4)性别影响调节子(sex-influenced modifier)和CNV也会在不同年龄表达形状，并且还会涉及如基因多效性等其他复杂因素。上述列举的结论是为了强调表型效应的广度，并且表明了当前被排除在外的单一突变(singleton variant)在后续有更多可供分析的样本加入后也许能重新被加入该CNV图谱中。

另外，更大更广泛的样本收集数据可以用于生成特定人群的CNV图谱（这有助于发现地理分离的突变(geographically segregated variants)），也能够用于生成仅包含罕见突变的图谱。特定人群的CNV图谱有助于研究CNV的正向或负向选择，或者在特定人群中与疾病或有利表型之间的关系。

CNV大小和基因组特征突变性之间的相互作用关系需要进一步地研究。举例来说，某些特定的基因组特征更倾向于形成包含其全长的大型重复突变，但是却因为很多小型重复突变改变了它们的结构。

我们还需要指出的是在当前图谱中每个CNVR中的拷贝数水平(copy number level)没有特意标注，因为很多本次使用到的数据集没有进行基因分型进而无从捕捉进一步的信息。随着算法改进和基因测序技术更迭，我们将获得一个更加全面的基因组突变的目录，用于准确地报道每个CNVR的人群频率。

DGV中经审核过的公开突变信息将进一步增加，并且该数据库已经成为用于科研和临床的主要CNV数据来源。DGV中收录的突变来自于健康人群，但是大量的可用表型信息仍然有限。举例来说，在一个癌症研究课题中健康对照人群不会进行血压或其他临床表征的评估；另外，健康不是固定不变的，受试者的身体状态会随时间发生变化。DGV同样不会排除迟发性疾病的突变，因此在进行相关健康结果预测时需要相当谨慎。相比之下，DECIPHER数据库的目标是收录具有高外显率且有明显表型的突变。在本课题组过去的研究中，他们曾发现在DGV和DECIPHER数据库中有健康人群和患病人群信息的重叠情况，而相关不正常的重叠即需要相关解释也需要更高分辨率CNVR、更加详细地表型信息和更加完善的外显现象解释理论。

最终结论，迄今最新人类基因组CNV图谱将为在演化、健康和疾病人群中基因组不平衡重排的理解提供参考。