如何解决遗传分析中的隐患基因结构注释不全

文章来源（http://www.360doc.com/content/19/0107/19/52645714_807291610.shtml）

背景知识

如何对基因组序列进行注释

基因组组装完成后，或者是完成了草图，就不可避免遇到一个问题，需要对基因组序列进行注释。注释之前首先得构建基因模型，有三种策略:

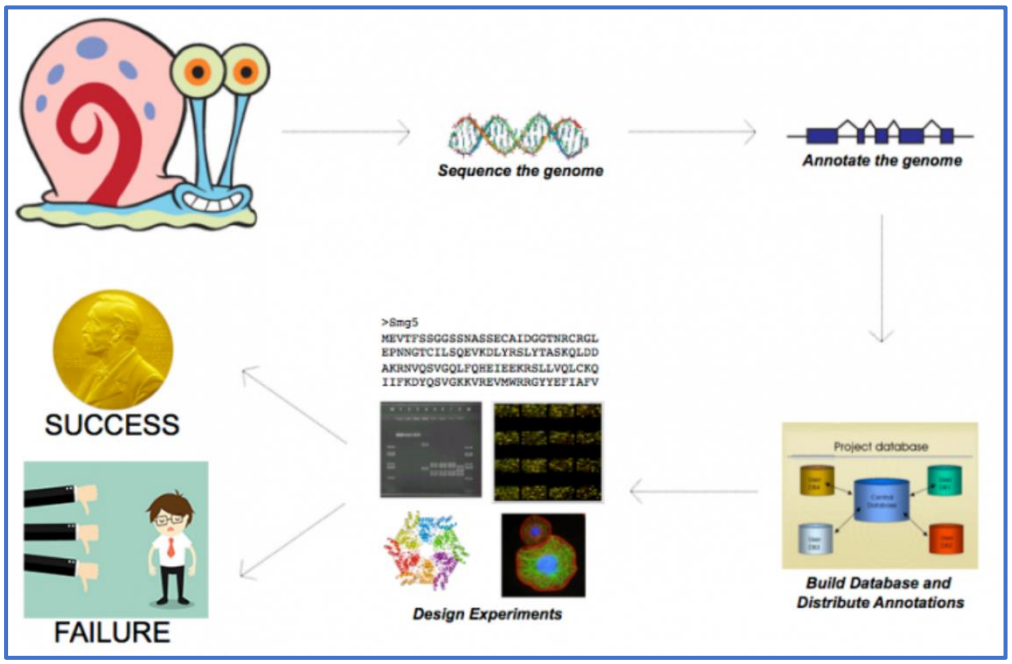
从头注释(de novo prediction): 通过已有的概率模型来预测基因结构，再预测剪切位点和 UTR 区，准确性较低。

同源预测(homology-based prediction): 有一些基因蛋白在相近物种间的保守性高，所以可以使用已有的高质量近缘物种注释信息通过序列联配的方式确定外显子边界和剪切位点。

基于转录组预测(transcriptome-based prediction): 通过物种的 RNA-seq 数据辅助注释，能够较为准确的确定剪切位点和外显子区域。

每一种方法都有自己的优缺点，所以最后需要用 EvidenceModelerEVM)GLEANT 工具进行整合，合并成完整的基因结构。基于可靠的基因结构，后续可才是功能注释，蛋白功能域注释，基因本体论注释，通路注释等。

那么基因注释重要吗？可以说是非常重要了，尤其是高通量测席非常便宜的现在。你可以花不到一万的价格对 600M 的物种进行 100X 的普通文库测序，然后拼接出草图。但是这个草图的价值还需要你进行注释后才能显现出来。有可能你和诺贝尔奖就差一个注释的基因组。（https://blog.csdn.net/u012110870/article/details/82500684）

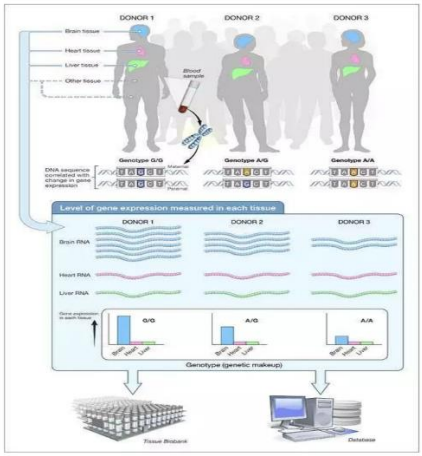


正文

当进行遗传分析时经常会遇到如下情况，OMIM 中明确报道基因未发现可疑变异，非编码区中发现一个罕见变异但无法分析，这两种情况通常会导致阴性或结果模糊的报告。

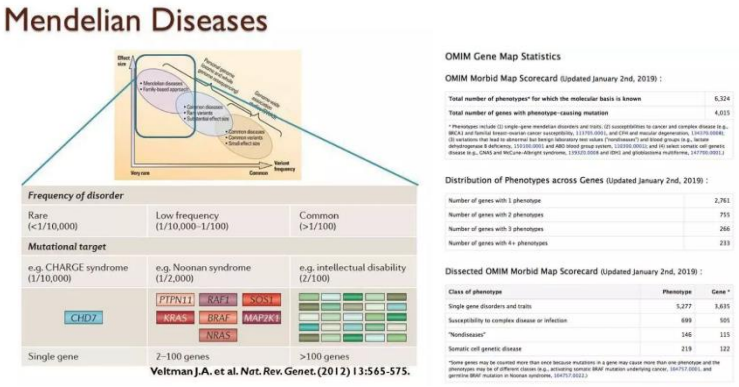
这两类问题的一个可能原因是基因结构注释不全，例如一些目前认为是不重要非编码区的部分其实有非常重要的生物学功能，而由于这些信息的缺失导致测序时未覆盖这些区域或者变异被标注为内含子变异。

个新的研究结果可能会给这两类问题带来解决的方向，这个研究的思路是利用 GTEx 的 RNA-seq 组学信息弥补缺失基因功能注释提供更多分析的证据。



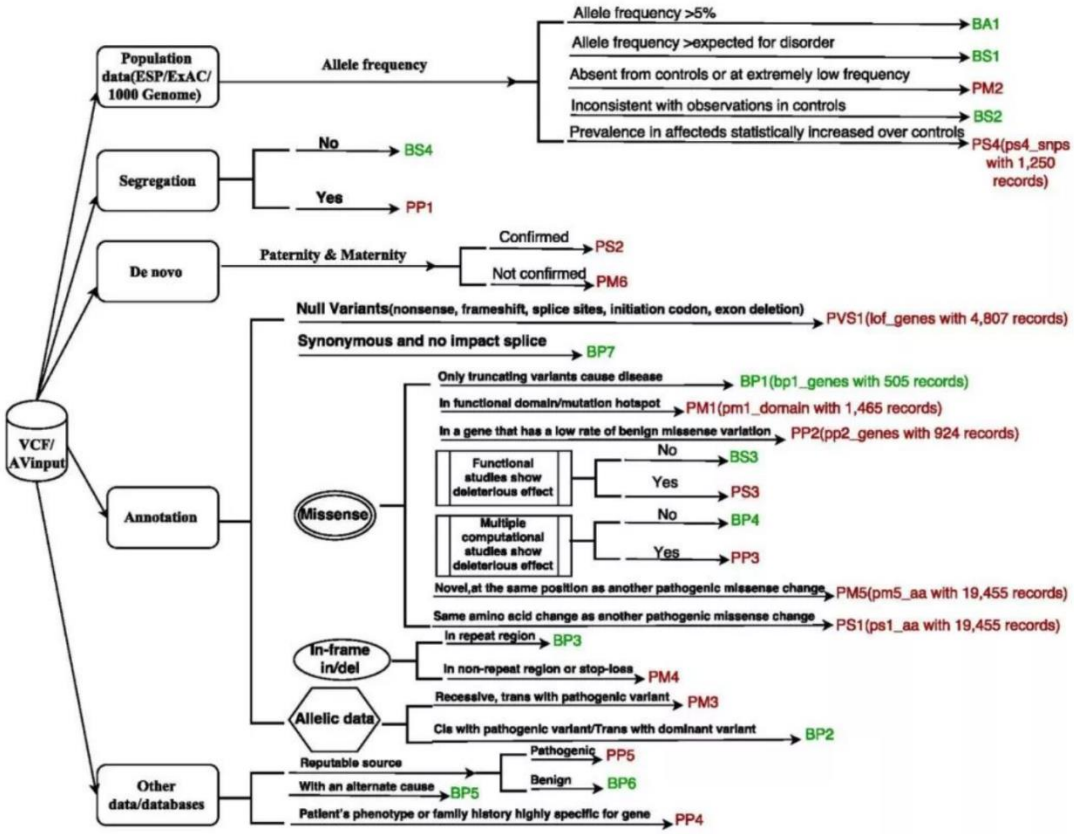
1 从 OMIM 数据库说起

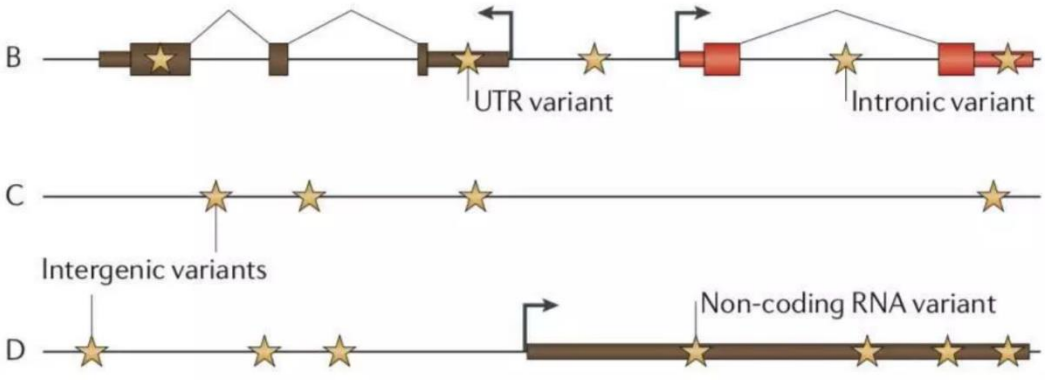
随着 NGS 技术的普及，越来越多的机构把 WES 甚至 WGS，当作遗传病分析的首选，孟德尔遗传疾病分析最常用的参考数据库非 OMIM 莫属，依靠其大量的专业人员维护，基因一疾病关系可信度非常高，是遗传分析证据的重要参考。



遗传检测的变异注释用 WES 或 WGS 做遗传病检测时，一个关键的步骤：分析结果中的变异性质。例如参考 ACMG 变异分类指南对检测到的变异进行致病性分类（例如 pathogenic），当目的基因上未发现候选致病变异时，就只能出非阳性报告。

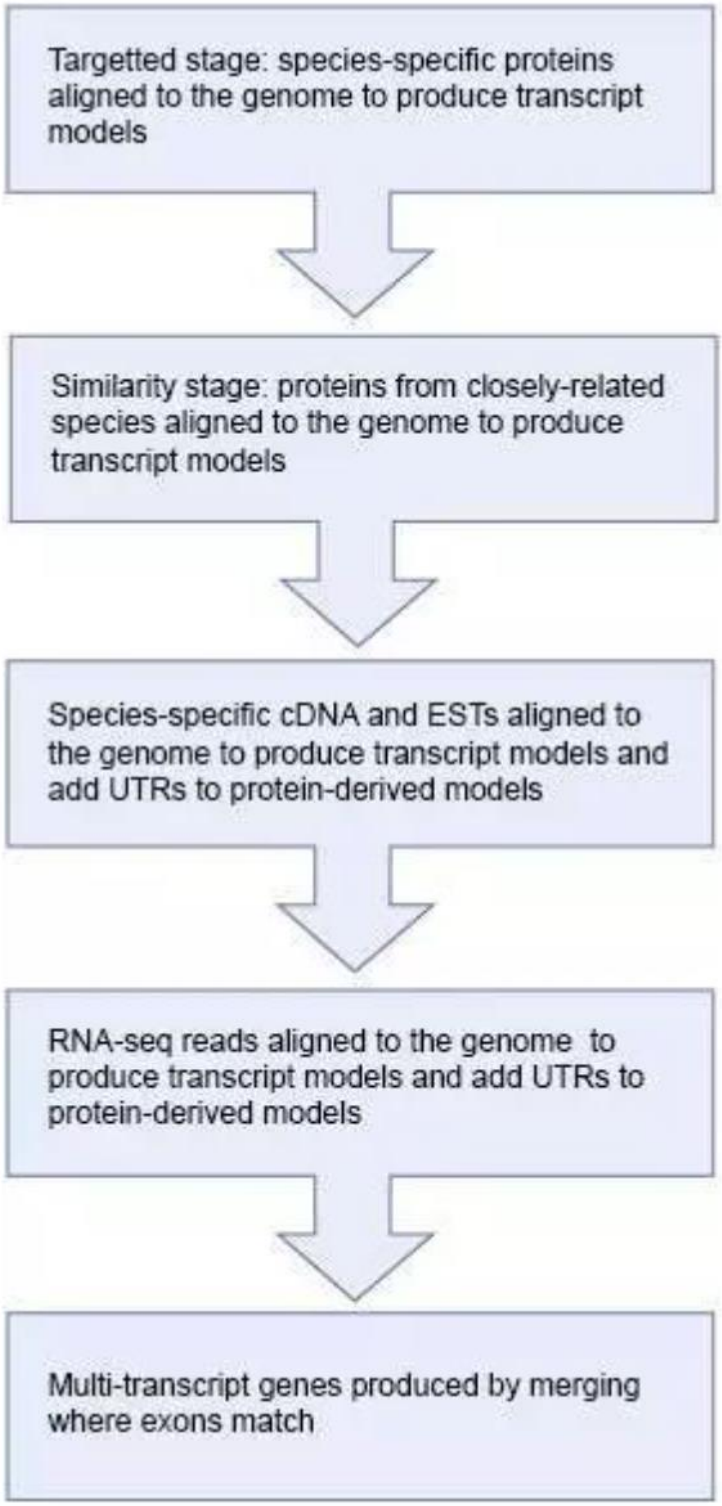
在这个过程中变异的注释非常依赖转录本参考数据库，例如 refseg 或 ensemb 这些变异对基因的影响，直接影响 ACMG 指南的使用。但 ACMG 指南对非编码区变异的证据支持很弱，所以非编码区的变异基本不在遗传分析的分析范围之内，这可能会漏掉许多非常重要但由于基因功能注释不全，导致无法分析的非编码区变异。

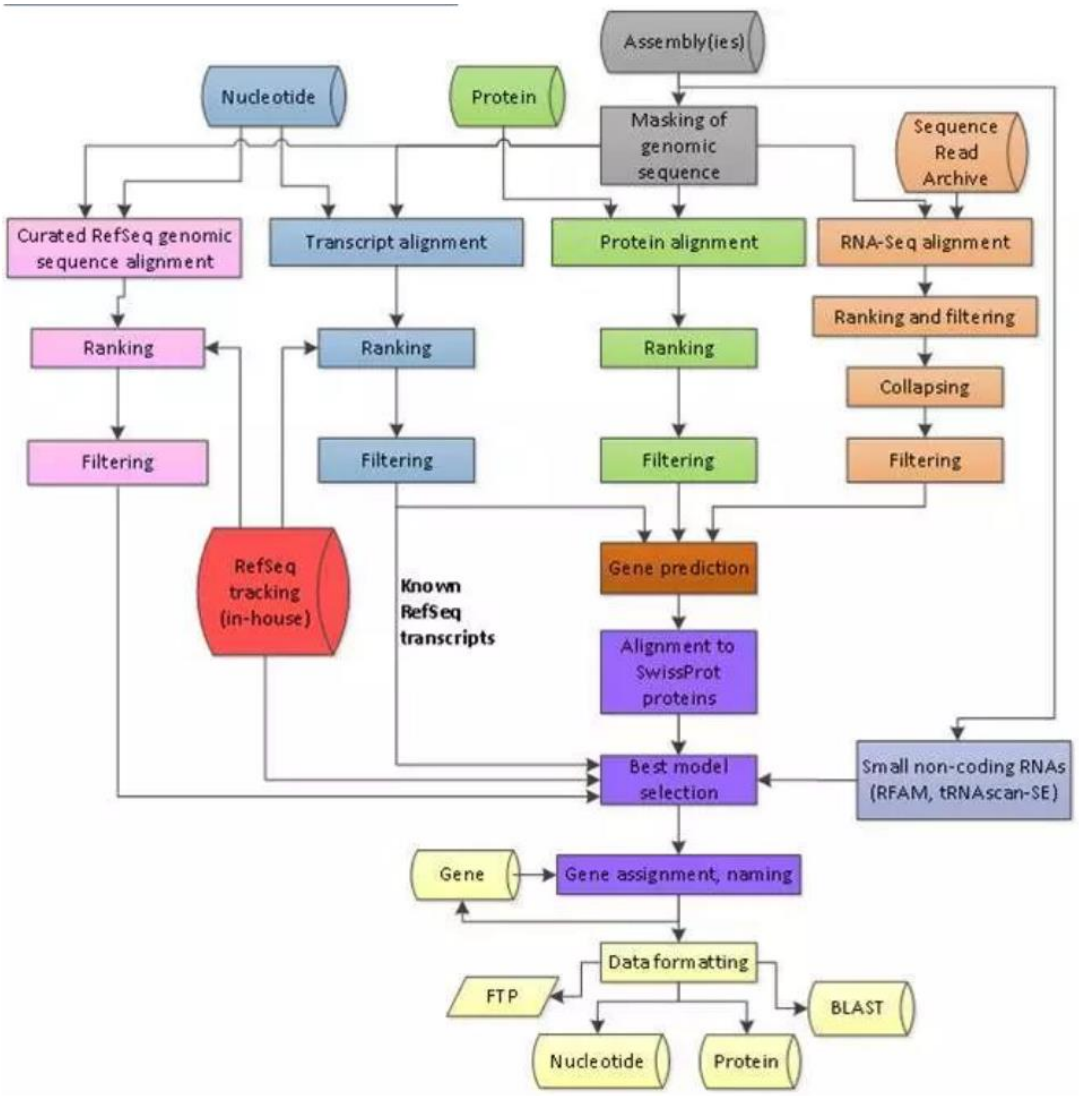

VCF/ AVinput
Population data(ESP/ExAC/ 1000 Genome)
Allele frequency
Allele frequency >5%
BA1
Allele frequency >expected for disorder
BS1
Absent from controls or at extremely low frequency
PM2
Inconsistent with observations in controls
BS2
Prevalence in affecteds statistically increased over controls
PS4(ps4_snps with 1,250 records)
Segregation
No
BS4
Yes
PP1
De novo
Paternity & Maternity
Confirmed
PS2
Not confirmed
PM6
Annotation
Null Variants(nonsense, frameshift, splice sites, initiation codon, exon deletion)
PVS1(lof_genes with 4,807 records)
Synonymous and no impact splice
BP7
Missense
Only truncating variants cause disease
BP1(bp1_genes with 505 records)
In functional domain/mutation hotspot
PM1(pm1_domain with 1,465 records)
In a gene that has a low rate of benign missense variation
PP2(pp2_genes with 924 records)
Functional studies show deleterious effect
No
BS3
Yes
PS3
Multiple computational studies show deleterious effect
No
BP4
Yes
PP3
Novel,at the same position as another pathogenic missense change
PM5(pm5_aa with 19,455 records)
Same amino acid change as another pathogenic missense change
PS1(ps1_aa with 19,455 records)
In-frame in/del
In repeat region
BP3
In non-repeat region or stop-loss
PM4
Allelic data
Recessive, trans with pathogenic variant
PM3
Cis with pathogenic variant/Trans with dominant variant
BP2
Other data/databases
Reputable source
Pathogenic
PP5
Benign
BP6
With an alternate cause
BP5
Patient's phenotype or family history highly specific for gene
PP4


B
UTR variant
Intronic variant
C
Intergenic variants
D
Non-coding RNA variant

2 如何解决注释不全

目前常用的基因注释数据库 refseg 与 ensemb 依赖常规数据库和生物信息学流程，对基因组进行基因标注例如下图中的注释流程 (ensembl 与 refseq 注释流程)。



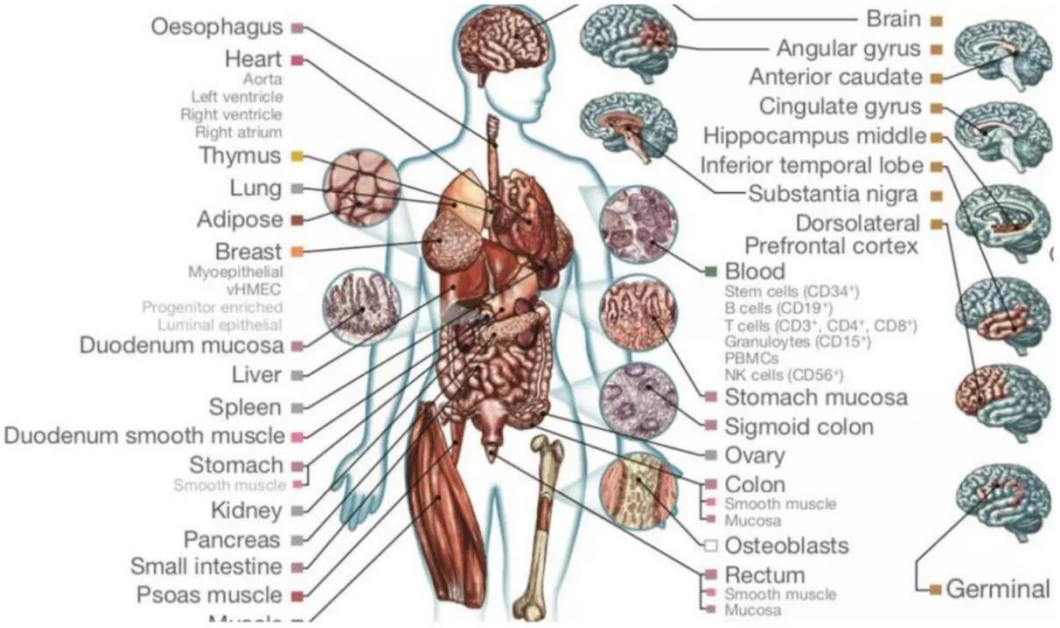


但目前的注释数据中，有许多可能遗漏的部分，**例如依赖同源基因的方法会漏掉人类特有的基因；依赖转录组数据的方法，会漏掉很多差异表达的基因**。

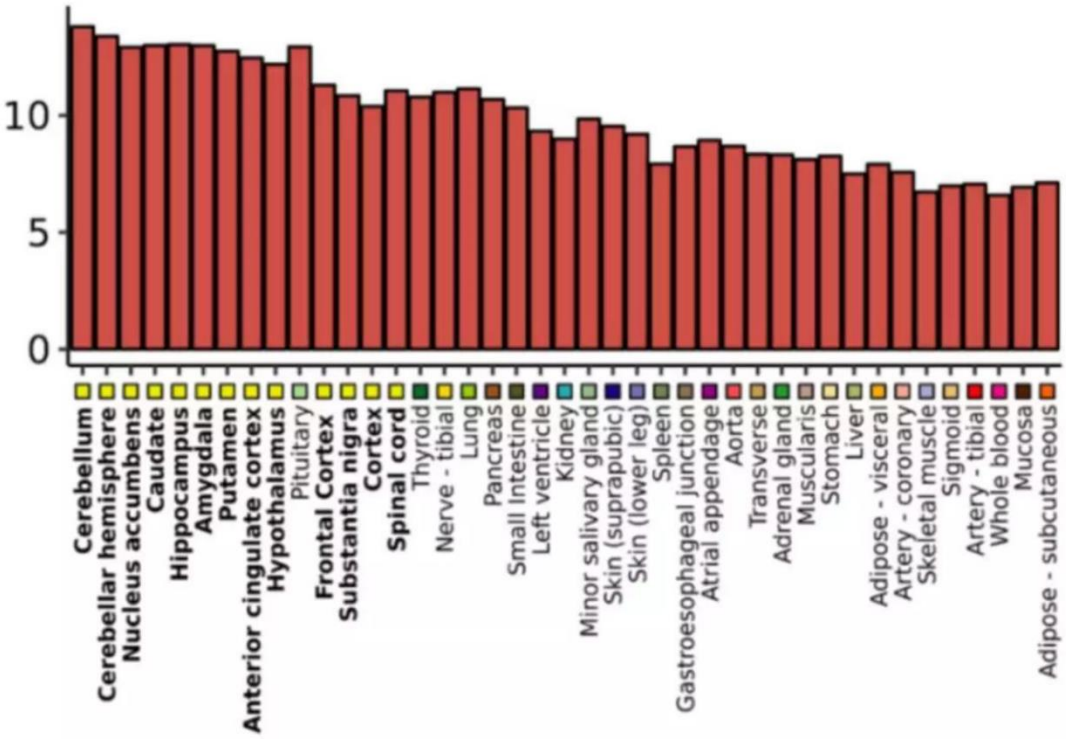
由于流程中存在的问题导致基因的功能注释不完整从而影响遗传病分析时的判断为了解决这个问题，一个来自多个研究机构的小组利用 GTEx 数据库中的数据弥补这些注释的空白。

3 GTEx 数据库是什么

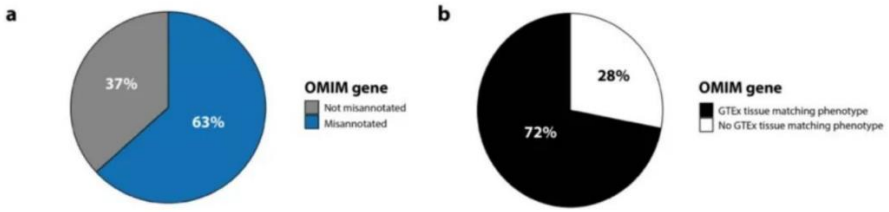
GTEx 数据库是一个 NIH 支持的疾病研究项目，主要目的是研究基因型与组织特异性基因表达的关系，研究变异在转录组水平的作用机制，因此数据库中有大量人类组织的转录组数据。



研究小组通过研究 41 种不同组织的 RNA 差异，发现有很多转录数据尤其是来自脑组织的 RNA 数据，在 refseq 或 ensembl 中未被注明，平均每个组织有 8.4M 区域在 ensembl v92 数据库中标记为非转录区域。

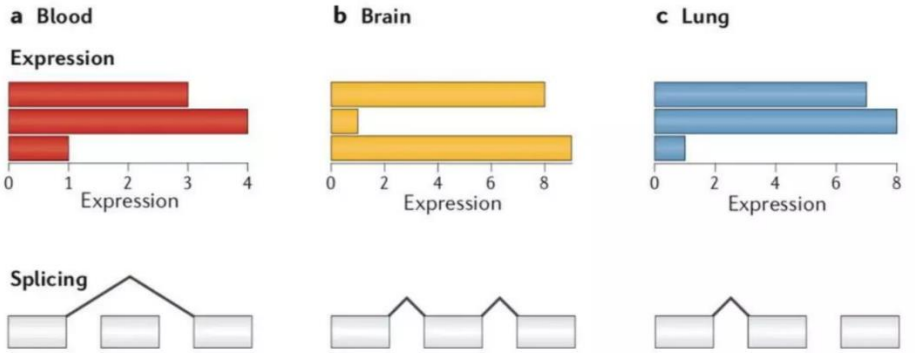


分析 RNA 数据中跨越标注区域与未标注区域的部分，这些未标注的区域通过参数优化和统计学检验，保留的数据被认为是有生物学功能但尚未标注的基因功能区域，这些标注差异大多数都是由于组织表达的差异性导致参考数据库的信息缺失，通过对 OMIM 数据库的分析，发现 63%的 OMIM 明确疾病相关基因存在基因功能标注不全，其中 72%的的标注不全对应疾病受累器官，脑组织的注释不全比例最大。



4 数据的潜在价值

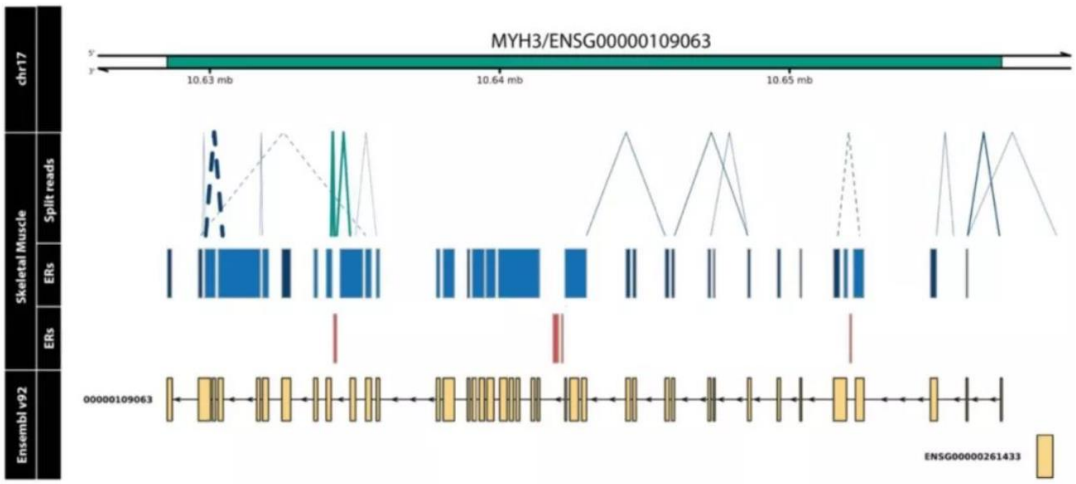
通过 GTEx 数据的补充，一些疾病的相关基因变异注释，可以变得更有参考价值。尤其是脑神经类疾病，由于其功能复杂性，组织特异转录现象非常普遍，而数据库中对特异转录标注非常缺乏，因此 GTEx 的挖掘数据非常适合这类疾病。对于其他组织的孟德尔疾病，GTEx 的挖掘数据也会有潜在作用。



文章中举了两个例子

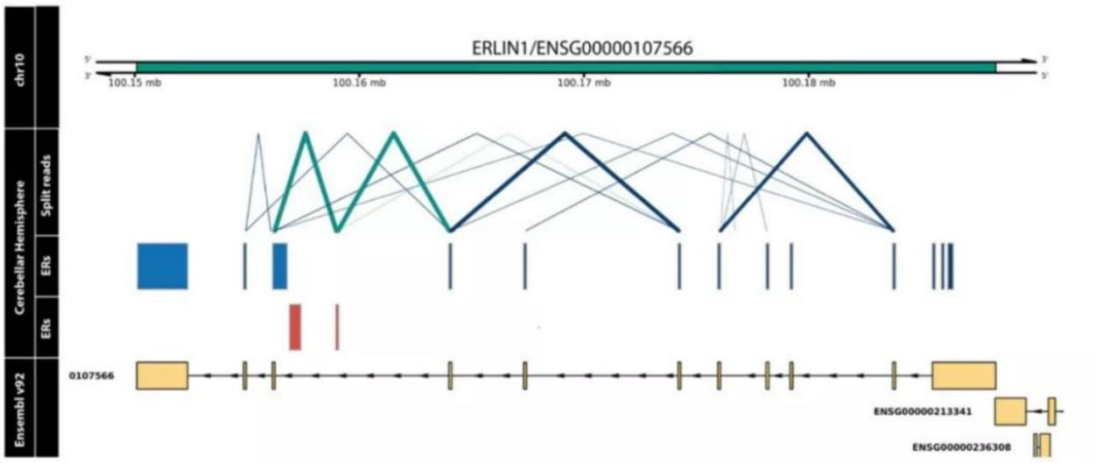
例 1:MYH3 基因

MYH3 基因会导致 distalarthrogryposis 的多种型，通过 GTEx 的数据分析，发现有一段 117bp，非保守但高限制的序列，只在肌肉组织中特异表达，但数据库中尚未标注，此段序列存在潜在的分析价值。



例 2:ERLIN1 基因

ERLIN1 基因会导致 spastic paraplegia 62 型此疾病的患者会有部分存在小脑症状，但原因尚未明确，通过 GTEx 数据库的分析发现一段 72bp 的序列，只在小脑中表达，但这段序列同样尚未标注。



5 总结

随着测序技术的发展，基因组序列也变得越来越完整，但对于基因组功能的标注一直是个高难度的问题，通过其他组学数据的辅助，如 RNA-seq，CHIP-seg 等会对基因组的功能注释提供很大帮助，当基因组功能注释更加完整时会提供更丰富参考信息，辅助遗传病分析的决策。