



# (12)发明专利申请

(10)申请公布号 CN 111540407 A

(43)申请公布日 2020.08.14

(21)申请号 202010288066.0

(22)申请日 2020.04.13

(71)申请人 中南大学湘雅医院

地址 410008 湖南省长沙市开福区湘雅路  
87号

(72)发明人 李津臣 李阔阔 赵贵虎 李滨

(74)专利代理机构 北京卓唐知识产权代理有限公司 11541

代理人 崔金

(51)Int.Cl.

G16B 30/00(2019.01)

G16B 40/00(2019.01)

G16H 50/20(2018.01)

G16H 50/70(2018.01)

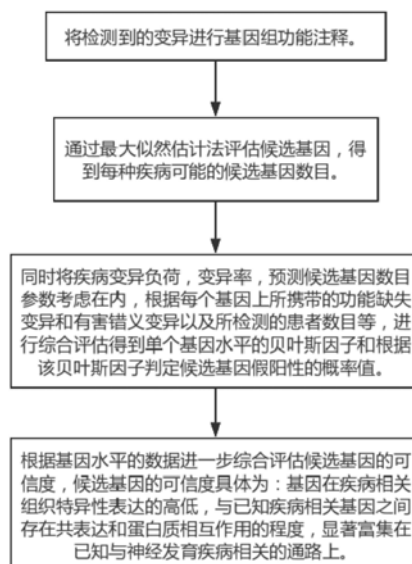
权利要求书2页 说明书5页 附图2页

## (54)发明名称

一种整合多种神经发育性疾病筛选候选基因的方法

## (57)摘要

本发明适用于神经疾病技术领域,提供了一种整合多种神经发育性疾病基因不同类别变异数据筛选候选基因的方法,基于最大似然估计的方法评估的某种疾病或某一类疾病的候选基因数目,基于贝叶斯模型方法同时结合期望基因变异率,病例变异负荷等数据鉴定神经疾病候选基因,具有整合多种神经发育性疾病多种类别变异数据的新思路,克服单种疾病和单种变异统计效应不足的问题,为鉴定新的候选基因、增强基因可信度提供了新方法。通过已知基因来验证本方法的可靠程度,可以根据疾病队列变异谱的分布情况寻找新的候选基因,同时该方法可以扩展到其他重大遗传疾病中。



1. 一种整合多种神经发育性疾病筛选候选基因的方法,其特征在于:包括以下步骤:

S1、通过最大似然估计法评估候选基因,得到每种疾病可能的候选基因数目;

S2、同时将疾病变异负荷,变异率,评估候选基因数目参数考虑在内,根据每个基因上所携带的功能缺失变异和有害错义变异以及所检测的患者数目等,进行综合评估得到单个基因水平的贝叶斯因子和根据该贝叶斯因子判定候选基因假阳性的概率值;

S3、根据基因水平的数据进一步综合评估候选基因的可信度。

2. 如权利要求1所述的一种整合多种神经发育性疾病筛选候选基因的方法,其特征在于:所述通过最大似然估计法评估候选基因,得到每种疾病可能的候选基因数目,具体为:

定义疾病所有功能缺失变异和有害错义变异的数目(K),同一个基因在不同患者中检测到两个及以上缺失变异和有害错义变异的基因数目(R)和功能缺失变异和有害错义变异对疾病有贡献的比例(E);

做一百万次置换检验,在每次置换检验中从所有编码基因中随机抽取一定比例的候选基因(T)(取值范围在1-2500)以及根据二项分布(K,E)抽取一定比例的变异数(C);

把对疾病有贡献的功能缺失变异和有害错义变异数(C)分配到候选基因(T)以及把对疾病无贡献的有害变异(K-C)分配给其他基因;

每次置换检验统计抽取到两次及以上基因的数目,如果这个数目和实际观测到的数目相同,就认为这次抽取的候选基因(T)为该疾病可能的致病基因数目;

所有的置换检验结束后,根据所有得到可能的致病基因数目得到一个先验概率分布曲线,概率最高的点所对应的候选基因数目为最可能是和疾病相关的基因数目。

3. 如权利要求1所述的一种整合多种神经发育性疾病筛选候选基因的方法,其特征在于:同时将疾病变异负荷、基因期望变异率、预测候选基因数目等参数考虑在内,根据每个基因上所携带的功能缺失变异和有害错义变异以及所检测的患者数目等,进行综合评估得到单个基因水平的贝叶斯因子和根据该贝叶斯因子判定候选基因假阳性的概率值,具体为:

观测在N个家系中出现的变异数目和期望的数目 $2N\mu$ , $\mu$ 表示特定基因的变异率,利用贝叶斯方法来比较 $H_0$ 和 $H_1$ 两个模型, $H_0$ 为基因不是致病基因, $H_1$ 为基因为致病基因;当 $H_0$ 成立,期望变异的数目为 $2N\mu$ ;当 $H_1$ 成立,期望变异的数目为 $2N\mu\gamma$ ,这里的 $\gamma$ 表示相对危险度, $\gamma$ 大于1表示可能是致病基因;

将贝叶斯因子定义为成为 $H_1$ 的概率除以 $H_0$ 的概率;根据贝叶斯定理确定 $H_1$ 为阳性的概率,贝叶斯因子大于1表示 $H_1$ 可能为阳性,贝叶斯因子大于100表示很强的证明 $H_1$ 为阳性;

将不同变异类型得到的贝叶斯因子相乘,每一种变异类型都会单独得到一个贝叶斯因子,把所有贝叶斯因子相乘,得到基因水平贝叶斯因子。

利用贝叶斯FDR的方法来控制假阳性率,每一个基因会得到一个FDR值,即q-value;

定义q-value小于一个预设值A,且 $0 \leq A \leq 1$ ,表示有A概率得到的基因可能是错误的。

4. 如权利要求1所述的一种整合多种神经发育性疾病筛选候选基因的方法,其特征在于:还包括:在执行步骤S1之前,将检测到的变异进行基因组功能注释,以将变异的类别分为功能缺失变异、有害错义变异、可耐受错义变异、同义变异和非编码区域变异。

5. 如权利要求1所述的一种整合多种神经发育性疾病筛选候选基因的方法,其特征在于:所述候选基因的可信度具体为:基因在疾病相关组织特异性表达的高低,与已知疾病相

关基因之间存在共表达和蛋白质相互作用的程度,显著富集在已知与神经发育疾病相关的通路上。

## 一种整合多种神经发育性疾病筛选候选基因的方法

### 技术领域

[0001] 本发明属于神经疾病领域,尤其涉及一种整合多种神经发育性疾病筛选候选基因的方法。

### 背景技术

[0002] 神经发育疾病由于其高的发病率和致残率,给个人和社会带来了很大的痛苦和负担。这类疾病种类繁多同时又表现出高度的临床异质性,给疾病的诊断和后续针对性的治疗带来了很大的困扰。目前对于疾病的诊断主要是根据病人或家属描述的表型和医生观察到的行为来判断。这些临床表型信息基本上都是根据以往专家总结的结果,并没有统一的生物学指标进行判断。

[0003] 神经发育疾病高的遗传度为研究疾病之间的相关性以及理解疾病的致病机制提供了很好的机会。以往的研究发现,在不同的神经发育疾病有很高的遗传相似度,并试图通过全基因组关联方法寻找可能的候选基因。但是这种方法找的候选基因数目非常少,紧能够解释很少的致病原因。由于不同种类神经疾病之间在临床表型和遗传机制上有很高概率的重叠,因此可以整合多种神经类疾病寻找可能的疾病候选基因。目前多个研究机构通过全外显子组或者全基因组测序方法检测新发变异(即只在患者中出现而在父母基因组中没有检测到的变异)并且能够找到许多候选基因。这种新发变异的很强功能破坏性,能够显著导致个体表型的改变。

[0004] 以往有方法通过整合不同的疾病变异数据寻找候选基因。但是这些方法目前有比较多的局限性,存在如下缺陷:

[0005] (1) 只考虑变异位点是否在已知可能候选基因上,这种方式虽然能够可能进行表型诊断,但是不能发现新的候选基因。

[0006] (2) 对于一个基因上出现多个可能有害变异,也不一定是候选基因。还需要考虑单个基因的变异率,所检测的病人数目,以及期望候选基因数目等参数。

[0007] (3) 以往的方法只考虑DNA水平变异和疾病的关系,并没有考虑组织特异性表达水平数据,蛋白水平等信息。所得到的证明可信度需要进一步分析。

[0008] (4) 没有真正考虑整合分析,以往方法只是单独疾病分析并做多个疾病候选比较,来研究不同疾病之间的遗传相关性。

### 发明内容

[0009] 本发明提供一种整合多种神经发育性疾病筛选候选基因的方法,旨在整合多种神经发育疾病新思路鉴定新的候选基因和增强基因可信度。

[0010] 本发明是这样实现的,一种整合多种神经发育性疾病筛选候选基因的方法,包括以下步骤:

[0011] S1、通过最大似然估计法评估候选基因,得到每种疾病可能的候选基因数目;

[0012] S2、同时将疾病变异负荷,变异率,评估候选基因数目参数考虑在内,根据每个基

因上所携带的功能缺失变异和有害错义变异以及所检测的患者数目等,进行综合评估得到单个基因水平的贝叶斯因子和根据该贝叶斯因子判定候选基因假阳性的概率值;

[0013] S3、根据基因水平的数据进一步综合评估候选基因的可信度。

[0014] 所述通过最大似然估计法评估候选基因,得到每种疾病可能的候选基因数目,具体为:

[0015] 定义疾病所有功能缺失变异和有害错义变异的数目(K),同一个基因在不同患者中检测到两个及以上缺失变异和有害错义变异的基因数目(R)和功能缺失变异和有害错义变异对疾病有贡献的比例(E);

[0016] 做一百万次置换检验,在每次置换检验中从所有编码基因中随机抽取一定比例的候选基因(T)(取值范围为1-2500)以及根据二项分布(K,E)抽取一定比例的变异数(C);

[0017] 把对疾病有贡献的功能缺失变异和有害错义变异数(C)分配到候选基因(T)以及把对疾病无贡献的有害变异(K-C)分配给其他基因;

[0018] 每次置换检验统计抽取到两次及以上基因的数目,如果这个数目和实际观测到的数目相同就认为这次抽取的候选基因(T)为该疾病可能的致病基因数目;

[0019] 所有的置换检验结束后,根据所有得到可能的致病基因数目得到一个先验概率分布曲线,概率最高的点所对应的候选基因数目为最可能是和疾病相关的基因数目。

[0020] 优选的,所述同时将疾病变异负荷,变异率,预测候选基因数目参数考虑在内,根据每个基因上所携带的功能缺失变异和有害错义变异以及所检测的患者数目等,进行综合评估得到单个基因水平的贝叶斯因子和根据该贝叶斯因子判定候选基因假阳性的概率值,具体为:

[0021] 观测在N个家系中出现的变异数目和期望的数目 $2N\mu$ , $\mu$ 表示特定基因的变异率,利用贝叶斯方法来比较 $H_0$ 和 $H_1$ 两个模型, $H_0$ 为基因不是致病基因, $H_1$ 为基因为致病基因;当 $H_0$ 成立,期望变异的数目为 $2N\mu$ ;当 $H_1$ 成立,期望变异的数目为 $2N\mu\gamma$ ,这里的 $\gamma$ 表示相对危险度, $\gamma$ 大于1表示可能是致病基因;

[0022] 将贝叶斯因子定义为成为 $H_1$ 的概率除以 $H_0$ 的概率;根据贝叶斯定理确定 $H_1$ 为阳性的概率,贝叶斯因子大于1表示 $H_1$ 可能为阳性,贝叶斯因子大于100表示很强的证明 $H_1$ 为阳性;

[0023] 将不同变异类型得到的贝叶斯因子相乘,每一种变异类型都会单独得到一个贝叶斯因子,把所有贝叶斯因子相乘,得到基因水平贝叶斯因子。

[0024] 利用贝叶斯FDR的方法来控制假阳性率,每一个基因会得到一个FDR值,即q-value;

[0025] 定义q-value小于一个预设值A,且 $0 \leq A \leq 1$ ,表示有A概率得到的基因可能是错误的。

[0026] 优选的,还包括:在执行步骤S1之前,将检测到的变异进行基因组功能注释,以将变异的类别分为功能缺失变异、有害错义变异、可耐受错义变异、同义变异和非编码区域变异。

[0027] 优选的,所述候选基因的可信度具体为:基因在疾病相关组织特异性表达的高低,与已知疾病相关基因之间存在共表达和蛋白质相互作用的程度,显著富集在已知与神经发育疾病相关的通路上。

[0028] 与现有技术相比,本发明的有益效果是:

[0029] 1、通过已有基因来验证本方法的可靠程度,另外可以根据未知基因有害变异来研究寻找新的候选基因。

[0030] 2、结合变异率,检测样本数目,候选基因数目等参数,从统计学上对候选基因的可信度进行评估。对发现的疾病致病基因有更可靠的依据。

[0031] 3、为了弥补单个神经发育疾病样本量不足的缺点,同时鉴定更多的候选基因。本方法基于不同神经发育疾病的遗传相似性,整合多个疾病的新发变异数据增强遗传统计学能力。

## 附图说明

[0032] 图1为本发明的实施例提供的一种整合多种神经发育性疾病筛选候选基因的方法的流程示意图;

[0033] 图2为本发明的一种整合多种神经发育性疾病筛选候选基因的方法的原理示意图。

## 具体实施方式

[0034] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合附图及实施例,对本发明进行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0035] 请参阅图1,本发明提供一种技术方案:一种整合多种神经发育性疾病筛选候选基因的方法,包括以下步骤:

[0036] S1、将检测到的变异进行基因组功能注释,以将变异的类别分为功能缺失变异、有害错义变异、无害错义变异、同义变异和非编码区域变异。功能缺失变异包括框移变异、剪切位点变异、终止变异和终止密码子缺失变异。从变异遗传方式上面考虑包括新发变异,传递和非传递变异和病例-对照变异。其中携带功能缺失变异和有害错义变异的基因更可能是候选基因。

[0037] S2、通过最大似然估计法评估候选基因,得到每种疾病可能的候选基因数目。并通过以下具体步骤实施:

[0038] 定义疾病所有可能功能缺失变异和有害错义变异的数目(K),同一个基因在不同患者中检测到两个及以上缺失变异和有害错义变异的基因数目(R)和功能缺失和有害错义变异对疾病有贡献的比例(E)。

[0039] 做一百万次置换检验,在每次置换检验中从所有编码基因中随机抽取一定比例的候选基因(T)(取值范围为1-2500)以及根据二项分布(K,E)抽取一定比例的变异数(C)。假设每种疾病可能致病的基因不会超过2500个。

[0040] 把对疾病有贡献的功能缺失和有害错义变异数(C)分配到候选基因(T)以及把对疾病无贡献的有害变异(K-C)分配给其他基因。

[0041] 每次置换检验统计抽取到两次及以上基因的数目,如果这个数目和实际观测到的数目相同就认为这次抽取的候选基因(T)为该疾病可能的致病基因数目。

[0042] 所有的置换检验结束后,根据所有得到可能的致病基因数目得到一个先验概率分

布曲线,概率最高的点所对应的候选基因数目为最可能是和疾病相关的基因数目。

[0043] S3、同时将疾病变异负荷,变异率,预测候选基因数目参数考虑在内,根据每个基因上所携带的功能缺失变异和有害错义变异以及所检测的患者数目等,进行综合评估得到单个基因水平的贝叶斯因子和根据该贝叶斯因子判定候选基因假阳性的概率值。并通过以下具体步骤实施:

[0044] 观测在N个家系中出现的变异数目和期望的数目 $2N\mu$ , $\mu$ 表示特定基因的变异率,首先利用贝叶斯方法来比较 $H_0$ 和 $H_1$ 两个模型, $H_0$ 为基因不是致病基因, $H_1$ 为基因为致病基因。当 $H_0$ 成立,期望变异的数目为 $2N\mu$ 。当 $H_1$ 成立,期望变异的数目为 $2N\mu\gamma$ ,这里的 $\gamma$ 表示相对危险度, $\gamma$ 大于1表示可能是致病基因。

[0045] 将贝叶斯因子定义为成为 $H_1$ 的概率除以 $H_0$ 的概率。根据贝叶斯定理确定 $H_1$ 为阳性的概率,贝叶斯因子大于1表示 $H_1$ 可能为阳性,贝叶斯因子大于100表示很强的证明 $H_1$ 为阳性。

[0046] 将不同变异类型得到的贝叶斯因子相乘,每一种变异类型都会单独得到一个贝叶斯因子,把所有贝叶斯因子相乘,得到基因水平贝叶斯因子。

[0047] 利用贝叶斯FDR的方法来控制假阳性率,每一个基因会得到一个FDR值,即q-value。

[0048] 定义q-value小于一个预设值A,且 $0 \leq A \leq 1$ ,表示有A概率得到的基因可能是错误的。

[0049] S4、根据基因水平的数据进一步综合评估候选基因的可信度,候选基因的可信度具体为:基因在疾病相关组织特异性表达的高低,与已知疾病相关基因之间存在共表达和蛋白质相互作用的程度,显著富集在已知与神经发育疾病相关的通路上。

[0050] 其中,一般认为,一个致病基因会有更高的变异负荷,即相同患者样本比随机情况下检测到更多的变异数目。这个可以简单的根据泊松检验分析得到结果。我们可以比较观测到的在N个家系中出现的变异数目和期望的数目 $2N\mu$ ,这里 $\mu$ 表示特定基因的变异率。但是这种方式的问题就是没有考虑每种变异类型的功效,比如功能缺失变异和错义变异对疾病的影响可能是不一样的。本方法检测每个基因的变异负荷的时候考虑了不同变异类型对疾病表型的贡献度大小。比如功能缺失变异为高风险的变异类型,将会给一个相对高的权重。

[0051] 对于不同类别的神经发育类疾病,本方法首先单独疾病分析鉴别候选基因。同时鉴于不同神经发育类疾病之间的遗传相似度,我们采用整合所有神经发育类新发变异的方法。这样可以鉴定新的在不同疾病中共有的,而在单个疾病中由于样本量不够无法鉴别的致病基因。

[0052] 上述鉴定的基因利用基因表达,蛋白质相互作用,富集分析等方法进一步增强基因可信度。基因是否表达在特定的与疾病相关的组织中,和已知基因有无共表达或者蛋白质相互作用。候选基因集合有无富集在已知的和疾病相关的基因集或者信号通路上。

[0053] 本发明的一种整合多种神经发育性疾病筛选候选基因的方法,具有整合多种神经发育疾病新思路鉴定新的候选基因和增强基因可信度。通过已有基因来验证本方法的可靠程度,另外可以根据未知基因的一致变异来研究寻找新的候选基因。结合变异率,检测样本数目,候选基因数目等参数,从统计学上对候选基因的可信度进行评估。对发现的疾病致病基因有更可靠的依据。为了弥补单个神经发育疾病样本量不足的缺点,同时鉴定更多的候

选基因。本方法基于不同神经发育疾病的遗传相似性,整合多个疾病的新发变异数据增强遗传统计学能力。

[0054] 以上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,均应包含在本发明的保护范围之内。

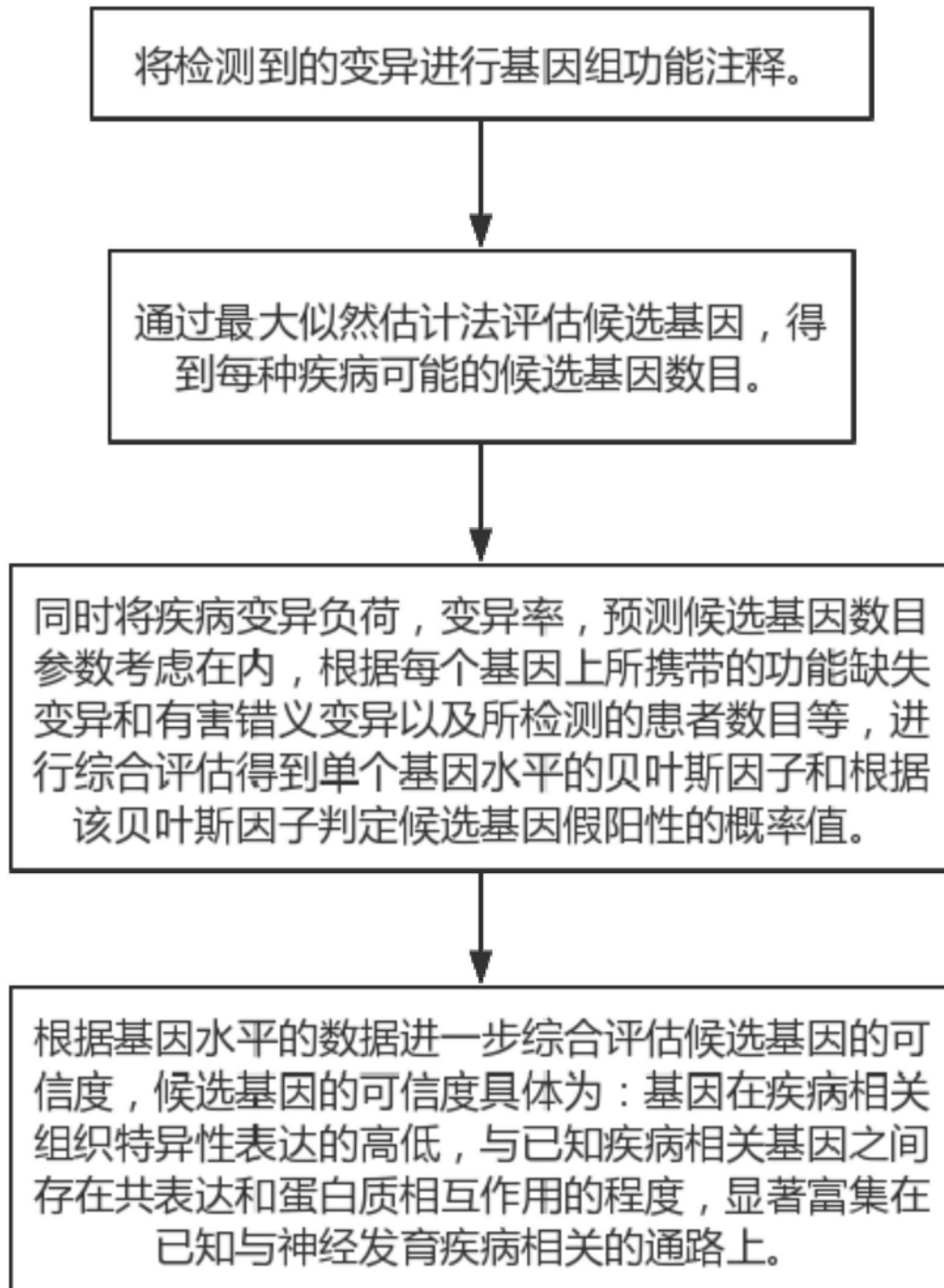


图1

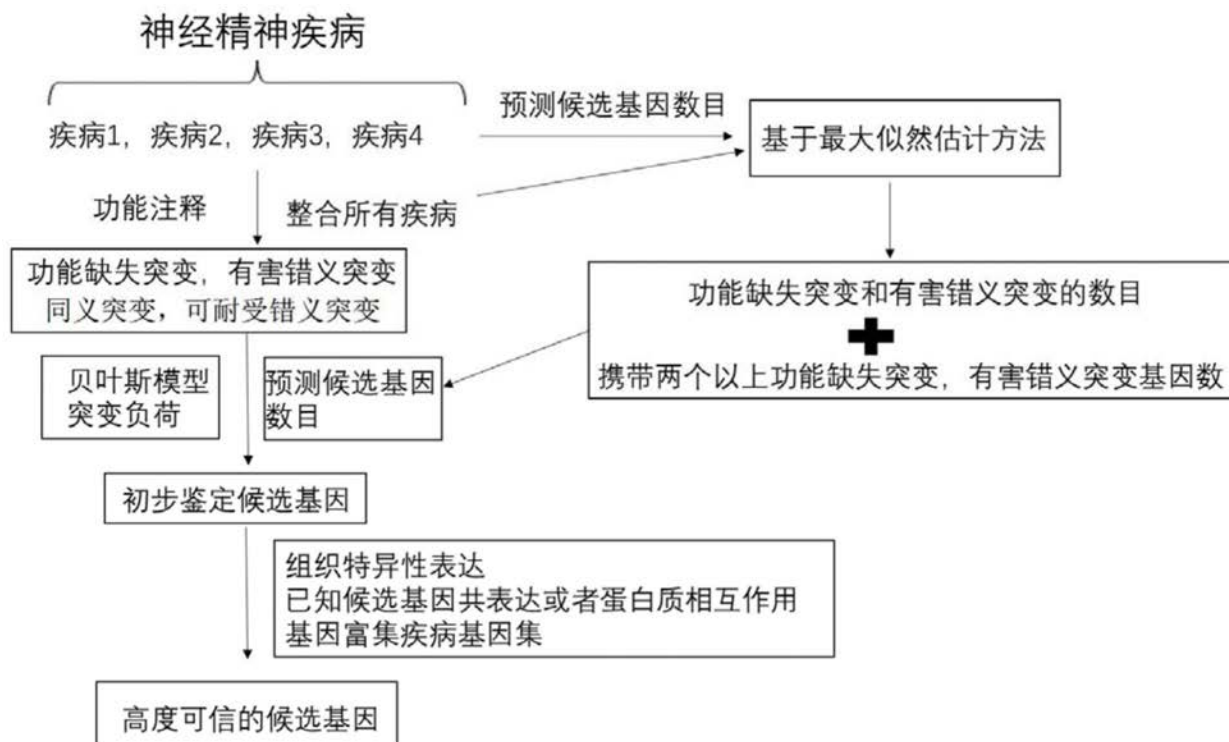


图2