

## 一种基因突变文库的构建方法

### 5 技术领域

本发明涉及基因工程领域，具体地涉及一种大库容基因突变文库的构建，本发明通过合成较少的寡聚物序列然后组装而得到大库容的基因突变文库。

### 10 背景技术

体外分子优化可非常有效地生成改良或新型的突变蛋白，识别调控序列，并可探测结构和功能的关键残基。用体外分子优化的方法构建合成文库对蛋白特性、调控和功能的系统性研究是非常有效的途径。应用高通量半导体精准引物池合成构建基因突变文库理论上所需的全部引物，能根据下游的不同表达系统做密码子优化，设置每个突变位置不同密码子分布率，从根本上杜绝非预期的密码子或终止密码子的引入问题，文库只包含所需要的突变体，为后续筛选节省的时间和精力。但是目前高通量半导体精准引物池合成的寡聚物数量有限，想要通过半导体芯片合成大库容的文库(10<sup>5</sup>-10<sup>10</sup>)成本非常高。

### 20 参考文献：

Kitzman J O, Starita L M, Lo R S 等, Massively parallel single-amino-acid mutagenesis[J]. *Nature Methods*, 2015, 12(3):203-206。

Kosuri S, Eroshenko N, Leproust E M 等, Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips[J]. *Nature Biotechnology*, 2010, 28(12):1295-1299。

### 发明内容

有鉴于此，发明了一种大库容文库的构建方法，能通过合成较少的寡聚物序列然后组装而得到大库容的基因突变文库，大幅度降低了大库容的文库构建的成本。

具体地，本发明涉及如下方面：

本发明的第一个方面涉及一种构建大库容基因突变文库的方法，所述方法包括：

(1) 根据需要构建文库的氨基酸序列的编码核苷酸序列设计并合成具有突变核苷酸和限制性核酸内切酶的酶切位点的两个或多个寡聚物池，其中设计的相邻的两个寡聚物池酶切后产生相同的粘性末端；

(2) 扩增所述寡聚物池；

(3) 在反应体系中组装寡聚物池，得到组装的寡聚物池；

(4) 扩增组装后的寡聚物池，得到大库容的基因突变文库。

在一个实施方案中，本发明的方法中的步骤(2)包括：分别以所述各寡聚物池作为模板，以根据所述各寡聚物池的序列设计的正向引物和反向引物作为引物对，使用高保真 DNA 聚合酶对所述各寡聚物池进行 PCR 扩增，得到扩增的各寡聚物池。优选地，所述 PCR 扩增体系还包含牛血清白蛋白。

在一个实施方案中，本发明的方法中的步骤步骤(3)包括：加入所述扩增的各寡聚物池，同时加入限制性核酸内切酶和 DNA 连接酶，使用边切边连的方式对所述各扩增的寡聚物池进行组装，得到组装后的寡聚物池。

在一个实施方案中，本发明的方法中的步骤步骤(4)包括：以组装后的寡聚物池作为模板，以所设计的第一个寡聚物池的正向引物和最后一个寡聚物池的反向引物作为引物对，使用高保真 DNA 聚合酶进行 PCR，对组装后的寡聚物池进行 PCR 扩增，从而得到大库容的基因突变文库。

在一个实施方案中，本发明的基因突变文库的库容为多达  $10^5$ ，优选多达  $10^6$ ，更优选多达  $10^7$ ，更优选多达  $10^8$ ，更优选多达  $10^9$ ，更优选多达  $10^{10}$ 。

在一个实施方案中，本发明使用的限制性核酸内切酶是 IIS 型限制性核酸内切酶，且相邻的两个寡聚物池酶切后产生相同的粘性末端。所述 IIS 型限制性核酸内切酶选自下组中的一种或多种：*AcuI*、*AlwI*、*BbsI*、*BbvI*、*BccI*、*BceAI*、*BciVI*、*BfuAI*、*BmrI*、*BpmI*、*BpuEI*、*BsaI*、*BseRI*、*BsgI*、*BsmAI*、*BsmBI*、*BsmFI*、*BspMI*、*BspQI*、*BsrDI*、*BtgZI*、*BtsCI*、*BtsI*、*EarI*、*EciI*、*EcoP15I*、*FauI*、*FokI*、*HgaI*、*HphI*、*HpyAV*、*MboII*、*MmeI*、*MnII*、*PleI*、*SapI* 和 *SfaNI*。优选地，所述 IIS 型限制性核酸内切酶是 *BsaI*。

在一个实施方案中，本发明的方法中的步骤(1)包括如下：

(i) 在所述编码核苷酸序列中确定粘性末端，以所述粘性末端 3'端将所述序列分成对应于所述两个或多个寡聚物池的两段或多段序列，

(ii) 如果将所述序列分成两段，则：

在第一段粘性末端后的 3'端依次引入限制性核酸内切酶的识别序列的反向互补序列和特异性序列 1，从而得到寡聚物池 1，且在第二段的 5'端依次引入粘性末端、限制性核酸内切酶的识别序列和特异性序列 2，从而得到寡聚物池 2；

如果将所述序列分成 n 段，n 为大于等于 3 的正整数，则：

在第一段粘性末端后的 3'端依次引入限制性核酸内切酶的识别序列的反向互补序列和特异性序列 1，从而得到寡聚物池 1，且在第二段的 5'端依次引入粘性末端、限制性核酸内切酶的识别序列和特异性序列 2，在第二段粘性末端后的 3'端依次引入限制性核酸内切酶的识别序列的反向互补序列和特异性序列 3，从而得到寡聚物池 2，以此类推，且在第 n 段的 5'端依次引入粘性末端、限制性核酸内切酶的识别序列和特异性序列 2n-2，在第二段粘性末端后的 3'端依次引入限制性核酸内切酶的识别序列的反向互补序列和特异性序列 2n-1，从而得到寡聚物池 n。

在一个实施方案中，所述特异性序列 1-特异性序列 2n-1 可为与原编码核苷酸序列不同源的随机序列。

在一个实施方案中，所述粘性末端为单粘性末端或简并粘性末端。在一个实施方案中，所述粘性末端为单粘性末端，且所述寡聚物池的个数为 2-6 个，优选 2 个。在另一个实施方案中，其中所述粘性末端为简并粘性末端，且所述寡聚物池的个数为 2 个。在一个实施方案中，所述粘性末端的 GC 含量为 50-75%。在一个实施方案中，所述粘性末端不含回文结构。

在一个实施方案中，本发明的方法中的步骤(3)的边切边连的方式是 Golden Gate 方式。

在一个实施方案中，本发明的方法进一步包括：(5) 对步骤(4)得到的基因突变文库产物进行回收和/或纯化，以得到最终文库产物。优选地，通过凝胶电泳对步骤(4)得到的基因突变文库产物进行回收和/或纯化。

在一个实施方案中，本发明的方法进一步包括：(6) 对步骤(5)中得到的最终文库产物进行测序，以验证序列分布和/或检测氨基酸分布。优选地，所述测序为 Sanger 测序和/或 NGS 测序。

在一个实施方案中，本发明的方法在步骤(1)中合成的每个寡聚物池中的突变核苷酸的个数为 1-108 个，优选 1-21 个。在一个实施方案中，本发

明的方法在步骤(1)中合成的每个寡聚物池中的突变核苷酸编码 1-36 个, 优选 1-7 个突变氨基酸。在一个实施方案中, 本发明的方法在步骤(1)中合成的每个寡聚物池中的突变核苷酸所编码的突变氨基酸是相邻或不相邻的。

在一个实施方案中, 本发明的方法中所使用的高保真 DNA 聚合酶选自  
5 下组中的一种或多种: Phusion DNA 聚合酶、Q5 聚合酶和 primerSTAR 聚合酶。在一个优选实施例, 所使用的高保真 DNA 聚合酶为 Phusion DNA 聚合酶。

本发明的第二个方面涉及使用根据本发明的方法构建的基因突变文库。

本发明的第三个方面涉及根据本发明的方法构建的基因突变文库在蛋  
10 白或多肽筛选中的用途。

本发明的第四个方面涉及一种解析蛋白质中的氨基酸突变与蛋白质的特性、调控和/或功能之间的关系的方法, 其包括以下步骤:

(1) 使用根据本发明的方法构建基因突变文库;

(2) 将所构建的基因突变文库中的突变基因编码的蛋白质的特征、调控  
15 和/或功能与未突变的蛋白质的情况进行比较; 和

(3) 解析蛋白质中的氨基酸突变与蛋白质的特性、调控和/或功能之间的关系。

#### 本发明解决的技术问题和实现的技术效果

突变文库对于蛋白质工程和抗体药工程非常重要, 利用本发明的方法构  
20 建突变文库能为后续筛选节省大量的时间、精力和金钱。目前现有技术(包括寡聚物池的高通量半导体精准合成)所合成的寡聚物数量有限, 想要合成大库容(库容为  $10^5$  以上)的基因突变的文库非常困难或者成本非常高。本发明公开了一种大库容文库的构建方法, 能通过合成较少的寡聚物序列然后组装而构建成大库容的基因突变文库。库容可高至  $10^7$ - $10^{10}$ , 大大的节约了大  
25 库容的文库构建的成本, 解决了合成大库容文库价格昂贵的问题。

#### **附图简述**

通过以下详细的描述并结合附图将更充分地理解本发明。

图 1 是用于设计寡聚物池的 Library generate 程序的界面;

30 图 2 是具有单粘性末端的寡聚物池 1 和池 2 的扩增产物的电泳图(泳道 1: DL3000; 泳道 2: 寡聚物池-1; 泳道 3: 寡聚物池-2);

图 3 是单粘性末端组装后的寡聚物池的峰图；

图 4 是单粘性末端组装后的寡聚物池的扩增产物的电泳图(泳道 1: 扩增后的寡聚物池；泳道 2: DL3000)；

图 5 是单粘性末端组装得到的文库序列的 Sanger 测序结果；

5 图 6 是单粘性末端组装得到的文库序列的 NGS 测序的序列分布图；

图 7 是单粘性末端组装得到的文库序列的 NGS 测序的氨基酸分布；

图 8 是简并粘性末端组装后的寡聚物池的扩增产物的电泳图(泳道 1: DL3000；泳道 2: 扩增后的寡聚物池)；

图 9 是简并粘性末端组装得到的文库序列的 Sanger 测序结果；

10 图 10 是简并粘性末端组装得到的文库序列的 NGS 测序的序列分布图；

图 11 是简并粘性末端组装得到的文库序列的 NGS 测序的氨基酸分布。

## 发明详述

如本说明书和所附权利要求书中使用的，单数形式“一个”，“一种”和  
15 “该/所述”包括复数提及物，除非上下文明确另有规定。如此，例如，提及“一个/种分子”任选包括两个/种或更多个/种此类分子的组合，诸如此类。

如本文中使用的，术语“约”指技术领域技术人员容易知道的相应数值的常规误差范围。本文中提及“约”某个数值或参数包括(并描述)涉及该数值或参数本身的实施方案。

20 理解的是，本文中描述的本发明的方面和实施方案包括“包含”，“由……组成”，和“基本上由……组成”的方面和实施方案。

术语“基因突变文库”是指大量 DNA 变体序列的组合，是基因合成、基因突变和/或定向进化研究的产物。基因突变文库已经越来越多地应用于研究领域，如高通量药物靶点筛选，蛋白质工程定向进化，合成多样性抗体  
25 库用来筛选高亲和力和特异性的变异抗体等。构建基因突变文库是利用定点突变、定向进化等技术模拟自然选择过程，改变原有蛋白的氨基酸序列，以期获得具有特定功能的突变蛋白。基因突变文库包括：丙氨酸扫描文库、简并突变文库、三联体文库、定点饱和突变文库、随机突变文库等。

术语“基因突变文库的库容”是指基因突变文库的容量大小，即，基  
30 因突变文库中所含的 DNA 变体序列的数目。换言之，基因突变文库的库容越大，该文库中所含的 DNA 变体序列的数目越多。

术语“寡聚物”一般指的是由较少的重复单元所组成的聚合物，其相对分子质量介于小分子和高分子之间。本发明的寡聚物是指含有一定数目的核苷酸的核苷酸低聚物，可与寡核苷酸互换使用。本发明的寡聚物中所含核苷酸的数目可为 2-200 个、3-150 个、4-100 个、4-50 个、4-30 个。术语

5 “寡聚物池”在本发明中指含有多种不同寡聚物的混合物。

术语“突变(的)核苷酸”是指在多核苷酸或寡核苷酸序列的一个或多个(例如几个)位置发生改变的核苷酸，所述改变选自下组：取代、缺失、插入。取代意指将占据某位置的核苷酸用不同的核苷酸替代；缺失意指去除占据某位置的核苷酸；而插入意指在邻接并紧接着占据某位置的核苷酸之后添加核苷酸。类似地，术语“突变(的)氨基酸”是指在蛋白质或多肽序列的一个或多个(例如几个)位置发生改变的氨基酸，所述改变选自下组：取代、缺失、插入。取代意指将占据某位置的氨基酸用不同的氨基酸替代；缺失意指去除占据某位置的氨基酸；而插入意指在邻接并紧接着占据某位置的氨基酸之后添加氨基酸。

15 术语“限制性核酸内切酶”是指识别特定的核苷酸序列，并在每条链中特定部位的两个核苷酸之间的磷酸二酯键进行切割的一类活性蛋白质。限制性核酸内切酶分布极广，几乎在所有细菌的属、种中都发现至少一种限制性核酸内切酶。常见的酶切位点包括但不限于：*AgeI*、*BamHI*、*BglII*、*EcoRI*、*EcoRV*、*FseI*、*HindIII*、*MauBI*、*MluI*、*NheI*、*NotI*、*PacI*、*PmeI*、*PstI*、*SacI*、*SacII*、*Sall*、*SmaI*、*SpeI*、*SfiI* 等。在 DNA 序列上引入酶切位点可便于对该 DNA 序列进行分子生物学操作。

根据限制性核酸内切酶的结构/功能特性、大小、所需的辅助因子以及作用方式，可将其分为如下类型：I 型限制性核酸内切酶分子量较大，反应过程中除需  $Mg^{2+}$  外，还需要 S-腺苷-L 甲硫氨酸、ATP 等。它既能催化宿主 DNA 的甲基化(即，修饰作用)，又催化非甲基化的 DNA 的水解(即，识别和酶切作用)。通常，其酶切位点距离识别序列可达数千个碱基之远；II 型限制性核酸内切酶分子量通常较小，反应只需  $Mg^{2+}$ 。它只催化非甲基化的 DNA 的水解，且在识别序列(特异性的短回文序列)上有特异性的酶切位点，因此在遗传工程上被广泛应用。III 型限制性核酸内切酶与第一型限制性核酸内切酶类似，同时具有修饰及识别酶切的作用，但它识别短的不对称序列，且酶切位点与识别序列的距离大约 20-30 个碱基对。此外，还有一种

IIS 型限制性核酸内切酶，它们识别连续的非对称序列，且酶切位点与识别序列的距离大约 1-6 个碱基对。IIS 型限制性核酸内切酶的大小居中，为约 400-650 个氨基酸。一般认为 IIS 型限制性核酸内切酶主要以单体的形式结合到 DNA 上，但与临近的酶结合成二聚体，协同切开 DNA 链。因此一些

5 IIS 型的酶在切割有多个识别序列的 DNA 分子时，活性可能更高。常见的 IIS 型限制性核酸内切酶包括但不限于 *AcuI*、*AlwI*、*BbsI*、*BbvI*、*BccI*、*BceAI*、*BciVI*、*BfuAI*、*BmrI*、*BpmI*、*BpuEI*、*BsaI*、*BseRI*、*BsgI*、*BsmAI*、*BsmBI*、*BsmFI*、*BspMI*、*BspQI*、*BsrDI*、*BtgZI*、*BtsCI*、*BtsI*、*EarI*、*EciI*、*EcoP15I*、*FauI*、*FokI*、*HgaI*、*HphI*、*HpyAV*、*MboII*、*MmeI*、*MnII*、*PleI*、

10 *SapI*、*SfaNI* 等。

术语“DNA 连接酶”是指能够在一条 DNA 链的 3'-OH 末端和另一条 DNA 链的 5'-P 末端二者之间生成磷酸二酯键，从而把两段相邻的 DNA 链相连的酶。连接酶的催化作用需要消耗 ATP。T4 DNA 连接酶是一种可连接 DNA-DNA、DNA-RNA、RNA-RNA 和双链 DNA 粘性末端或平末端的

15 连接酶，分子量为约 62 kD，其活性易被 0.2 mol/L 的 KCl 和精胺所抑制。

术语“边切边连”是指在一次反应中可将多个片段按照指定的先后顺序相互连接的组装方法。首先在待组装片段上构造互补的粘性末端，然后将待组装的片段与限制性核酸内切酶(特别是 IIS 型限制性核酸内切酶)与 DNA 连接酶混合，分别在高温(限制性核酸内切酶的最适温度，例如 37°C)和低温(连接酶的最适温度，例如 16°C)之间循环多次。高温有利酶切而低温有利连接，每次循环会导致未连接或连接回原质粒的片段再次遭到酶切，从而以更大的概率保留正确组装的质粒。最后将两种酶失活。Golden Gate 公司首次使用 IIS 型限制性核酸内切酶 *BsaI* 和连接酶进行的边切边连的组装方法被称为 Golden Gate 方式，具体操作参见例如 Engler C 等, PLoS ONE, 2008, 3(11): e3647 和 Engler C 等, PLoS ONE, 2009, 4(5): e5553。

20 25

术语“引入”是指将外源性多核苷酸序列插入到 DNA 链中。插入后的外源性多核苷酸序列通常可以被正常表达。

术语“表达”是指信息(例如基因编码的和/或表观遗传的)转化成细胞中存在并运行的结构的过程。因此，如本文中使用的，“表达”可以指转录成多核苷酸，翻译成多肽，或甚至多核苷酸和/或多肽修饰(例如多肽的翻译后修饰)。转录的多核苷酸的片段、翻译的多肽的片段、或多核苷酸和/或

30

多肽修饰(例如多肽的翻译后修饰)的片段也应视为表达的, 无论它们是源自通过可变剪接生成的转录物或经过降解的转录物, 或者是源自多肽的翻译后加工(例如通过蛋白质水解)。“表达的基因”包括转录成多核苷酸(如 mRNA), 然后翻译成多肽的基因, 还有转录成 RNA 但不翻译成多肽的基因  
5 (例如转运和核糖体 RNA)。

术语“PCR 扩增”是指聚合酶链式反应, 其为用于扩增特定的 DNA 片段的分子生物学技术, 它可看作是生物体外的特殊 DNA 复制, PCR 的最大特点是能将微量的 DNA 大幅增加。具体而言, PCR 技术的基本原理类似于 DNA 的天然复制过程, 其特异性依赖于与待扩增序列的两端互补的寡核苷酸引物。术语“PCR 模板”是指待扩增的特定 DNA 片段。术语“PCR 引物(对)”是指一对寡核苷酸序列, 其中正向引物(F)是与作为模板的双链 DNA 中的有义链的 5'端碱基互补(A 对 T, C 对 G, 反之亦然)的寡核苷酸序列, 而反向引物(R)则是指与作为模板的双链 DNA 中的反义链的 3'端碱基互补(A 对 T, C 对 G, 反之亦然)的寡核苷酸序列。PCR 扩增由变性-退火-延伸三个基本反应  
10 步骤构成: ①模板 DNA 的变性: 模板 DNA 经加热至 90℃ 以上一定时间后, 使模板 DNA 双链或经 PCR 扩增形成的双链 DNA 解离, 使之成为单链, 以便它与引物结合, 为下轮反应作准备; ②模板 DNA 与引物的退火(复性): 模板 DNA 经加热变性成单链后, 温度降至约 50℃, 引物与模板 DNA 单链的互补序列配对结合; ③引物的延伸: DNA 模板-引物结合物在约 70℃ 在 DNA 聚合酶的作用下, 以 dNTP 为反应原料, 以模板为基础, 按碱基互补配对与半保留复制原理, 合成一条新的与模板 DNA 链互补的半保留复制链, 重复循环  
15 变性-退火-延伸三过程就可获得更多的“半保留复制链”, 而且这种新链又可成为下次循环的模板。术语“DNA 聚合酶”是指以亲代 DNA 为模板, 催化底物 dNTP 分子聚合形成子代 DNA 的一类酶。DNA 聚合酶的共同特征为:  
20 ①具有 5'→3'聚合酶活性, 这就决定了 DNA 只能沿着 5'→3'方向合成; ②由于 DNA 聚合酶不能催化 DNA 新链从头合成, 只能催化 dNTP 加入核苷酸链的 3'-OH 末端, 因此需要 DNA 引物作为合成的起点。高保真 DNA 聚合酶是指具有较高保真度的 DNA 聚合酶, 使用它合成的 DNA 与模板的同一性比其他 DNA 聚合酶更高。本发明可使用任何本领域已知的高保真 DNA 聚合酶,  
25 例如 Phusion DNA 聚合酶、Q5 聚合酶和 primerSTAR 聚合酶等。  
30

在构建本发明的基因突变文库而实施各种分子生物学操作时, 所采用

的规程和方法(包括但不限于合成、酶切、连接、PCR、回收、纯化等等)都是本领域技术人员所熟知的(参见,例如, Sambrook等, 1989, 分子克隆实验指南(第二版), 冷泉港, 纽约)。

5 本发明提供了一种构建大库容基因突变文库的方法, 解决了现有技术中大库容基因突变文库的构建成本奇高的难题。本发明所构建的基因突变文库的库容可为多达 $10^5$ , 优选多达 $10^6$ , 更优选多达 $10^7$ , 更优选多达 $10^8$ , 更优选多达 $10^9$ , 更优选多达 $10^{10}$ 。在一个具体实施方案中, 基因突变文库的库容为 $1.25 \times 10^7$ 。

10 提供需要构建文库的氨基酸序列并确定其中需要突变的氨基酸的位置和种类, 通过计算获得库容, 并确定编码所述氨基酸序列的核苷酸序列。在所述编码核苷酸序列中确定粘性末端的序列和位置, 以该粘性末端将编码核苷酸序列分成对应于所述两个或多个(如3-10个, 优选3-6个)寡聚物池的两段或多段(如3-10段, 优选3-6段)序列。粘性末端依据其序列的个数分为单粘性末端(其只有一种序列)和可将连续突变的点分成两段的简并粘性末端(其有很多  
15 种序列, 具体序列的数目根据粘性末端处的突变决定)。单粘性末端适用于将编码核苷酸序列分成两段以上, 如两段、三段、四段、五段、六段、七段、八段、九段、十段。粘性末端的序列选择基于如下标准: 1)粘性末端处于编码核苷酸序列中部, 能将寡聚物池序列作为长度大致相同两段分别合成; 2)粘性末端的序列中无回文序列; 3)粘性末端序列的GC含量为  
20 50%~75%; 4)粘性末端的长度与选择的IIS型限制性核酸内切酶的酶切位点相匹配。粘性末端的选择也可参考文献, 如Vladimir P, Ong J L, Kucera R B等, Optimization of Golden Gate assembly through application of ligation sequence-dependent fidelity and bias profiling[J]. BioRxiv, 2018: 322297。以适当的/错配率低的粘性末端将各寡聚物池组装起来, 然后用第一个寡聚物池  
25 的正向引物和最后一个寡聚物池的反向引物扩增得到最终片段。

本发明合成的寡聚物池中所含的限制性核酸内切酶的酶切位点是 IIS 型限制性核酸内切酶的酶切位点, 其通常距识别序列 1-6 bp, 优选 1-4bp。所述 IIS 型限制性核酸内切酶选自下组: *AcuI*、*AlwI*、*BaeI*、*BbsI*、*BbvI*、*BccI*、*BceAI*、*BcgI*、*BciVI*、*BfuAI*、*BmrI*、*BpmI*、*BpuEI*、*BsaI*、*BsaXI*、  
30 *BseRI*、*BsgI*、*BsmAI*、*BsmBI*、*BsmFI*、*BspMI*、*BspQI*、*BsrDI*、*BtgZI*、*BtsCI*、*BtsI*、*EarI*、*Ecil*、*EcoP15I*、*FauI*、*FokI*、*HaeIII*、*HgaI*、*HphI*、

*HpyAV*、*MboII*、*MlyI*、*MmeI*、*MnII*、*NmeAIII*、*PleI*、*SapI*、*SfaNI*。优选地，所述 IIS 型限制性核酸内切酶是 *BsaI*，其识别序列为 GGTCTC。

下面以分两段为例进一步详述本发明的方法：

将编码核苷酸序列分为两段之后，在第一段DNA序列的3'端加入IIS型  
5 限制性核酸内切酶(例如*BsaI*)的识别序列的反向互补序列，使得酶切后可保留上述确定的粘性末端，之后在引入识别序列后的DNA序列的3'端添加15-30bp，优选15-25bp，更优选18-21 bp的特异性序列，以便于后续扩增；在第二段DNA序列的5'端依次加入IIS型限制性核酸内切酶(例如*BsaI*)的识别序列和上述确定的粘性末端的序列，使得酶切后可保留与第一段DNA序列相  
10 同的粘性末端，在引入识别序列和粘性末端后的DNA序列的5'端添加15-30bp，优选15-25bp，更优选18-21 bp的特异性序列，以便于后续扩增。所述特异性序列可为与原编码核苷酸序列不同源的随机序列。

可使用本领域已知的任何方法来设计并合成第一段和第二段的寡聚物序列，例如，可使用Library generate程序来进行，可按照比例和合成引物的  
15 总条数计算第一段和第二段序列的条数。由于合成的寡聚物量太少，不能满足后续操作需求，因此需要对第一段寡聚物和第二段寡聚物进行扩增。通过本领域已知的方法设计寡聚物池-1和寡聚物池-2的扩增引物，寡聚物池-1的正向引物命名为1F，反向引物命名为1R。寡聚物池-2的正向引物命名为2F，反向引物命名为2R。所述引物的长度可为15-30bp，优选15-25bp，  
20 更优选18-21 bp。以合成的寡聚物池-1为模板、以1F/1R为引物，或者以寡聚物池-2为模板、以2F/2R为引物，加入高保真DNA聚合酶(例如Phusion DNA聚合酶)、dNTP、PCR缓冲液和任选的牛血清白蛋白(BSA)，分别对合成的寡聚物池进行扩增，以引物1F/1R得到扩增后的寡聚物池-1，以引物2F/2R得到扩增后的寡聚物池-2。扩增完成后通过凝胶电泳检测条带是否单  
25 一，若条带单一，则对扩增产物进行柱纯化；若扩增条带不单一，则需改变PCR条件重新扩增。所需改变的PCR条件包括：模板与引物的比例、退火温度、退火时间、循环次数等等。

得到纯化的扩增产物后，使用边切边连的方式对两段寡聚物进行组装，所述边切边连优选使用Golden Gate方式进行。配置组装体系，在体系  
30 中加入纯化的第一段寡聚物和第二段寡聚物，同时加入限制性核酸内切酶和DNA连接酶(和/或其缓冲液)。组装反应的程序是37℃、3分钟和16℃、5

- 分钟，以上两个步骤循环20次，37℃是限制性核酸内切酶的最适反应温度，16℃是DNA连接酶的最适反应温度。在组装过程中，先37℃反应3分钟，限制性核酸内切酶识别并切割酶切位点，产生上述确定的粘性末端，然后16℃反应5分钟，DNA连接酶将粘性末端序列相同的两段寡聚物进行连接。连接正确的片段上不含限制性核酸内切酶的识别序列而连接错误的片段含有限制性核酸内切酶的识别序列。在下一个循环至37℃时，限制性核酸内切酶可将连接错误的片段再次切开，直至片段连接正确。最后80℃反应15分钟，使两种酶失活，以免对后续反应造成影响。组装体系中的DNA量很低，纯化后基本得不到DNA，且可能含有未组装的片段，因此需要对组装后的序列进行PCR扩增。

- 以组装后的未纯化的产物为模板，加入第一段寡聚物的正向引物1F和第二段寡聚物的反向引物2R，用高保真DNA聚合酶(如Phusion DNA聚合酶等)进行PCR，对组装后的序列池进行扩增得到大库容的基因突变文库。之后可对PCR产物进行回收和/或纯化而得到终产物，所述回收和/或纯化优选通过凝胶电泳进行。将终产物进行平末端连接，将其连接线性化质粒(如Puc57-EV)，使用通用引物(如M13F引物)对终产物进行Sanger测序，检测序列是否正确；同时将终产物进行高通量测序技术(NGS)，检测序列分布和氨基酸分布等。

## 20 序列说明

| SEQ ID NO | 具体氨基酸或序列                                                                                                          |
|-----------|-------------------------------------------------------------------------------------------------------------------|
| 1         | RPDVNASX1X2X3GX4TPLHLAAX5X6GHLEIVEVLLKX7GADV RPD                                                                  |
| 2         | AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTGGCTGCTX5X6GGTCACCTGGAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC |
| 3         | AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTG                                                                   |
| 4         | GAGACC                                                                                                            |
| 5         | AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTGAGAGACC                                                            |
| 6         | ACCTACGTCCAAATCGGCTG                                                                                              |
| 7         | AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTGAGAGACCACCTACGTCCAAATCGGCTG                                        |
| 8         | GCTGCTX5X6GGTCACCTGGAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC                                                |
| 9         | GGTCTC                                                                                                            |
| 10        | GGTCTCAGCTGGCTGCTX5X6GGTCACCTGGAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC                                     |

|    |                                                                                                                                |
|----|--------------------------------------------------------------------------------------------------------------------------------|
| 11 | ACTCCTAAGTAAATGTGCAC                                                                                                           |
| 12 | ACTCCTAAGTAAATGTGCACGGTCTC <b>ACCT</b> GGCTGCTX5X6GGTCACCTGGAAATCGTTGAAGTTC<br>TGCTGAAGX7GGTGTGACGTGAGACCTGAC                  |
| 13 | 1F: AGACCTGACGTAAACGCTAG                                                                                                       |
| 14 | 1R: CAGCCGATTGGACGTAGGT                                                                                                        |
| 15 | 2F: ACTCCTAAGTAAATGTGCA                                                                                                        |
| 16 | 2R: GTCAGGTCTCACGTGAGCAC                                                                                                       |
| 17 | AGACCTGACGTAAACGCTAGCX1X2X3GGTX4ACTCCGCTGCAC <b>CT</b> GGCTGCTX5X6GGTCACCTGG<br>AAATCGTTGAAGTCTGTGCTGAAGX7GGTGTGACGTGAGACCTGAC |
| 18 | KTEDTAVYYCSRDI02103104105106107DAWGQGLTVTVSS                                                                                   |
| 19 | AAAACCGAGGACACGGCCGTGTATTACTGTAGTAGAGAT102103104105106107GACGCCTGGG<br>GCCAAGGAACCTGGTCACCGTCTCCTCG                            |
| 20 | AAAACCGAGGACACGGCCGTGTATTACTGTAGTAGAGAT102103104105                                                                            |
| 21 | AAAACCGAGGACACGGCCGTGTATTACTGTAGTAGAGAT102103104105AGAGACC                                                                     |
| 22 | AAAACCGAGGACACGGCCGTGTATTACTGTAGTAGAGAT102103104105AGAGACC<br>ACCTACGTCCAAATCGGCTG                                             |
| 23 | 104105106107GACGCCTGGGGCCAAGGAACCTGGTCACCGTCTCCTCG                                                                             |
| 24 | GGTCTCG104'105106107GACGCCTGGGGCCAAGGAACCTGGTCACCGTCTCCTCG                                                                     |
| 25 | ACTCCTAAGTAAATGTGCACGGTCTCG104'105106107GACGCCTGGGGCCAAGGAACCTGGT<br>CACCGTCTCCTCG                                             |
| 26 | 1sF: AAAACCGAGGACACGGCCGT                                                                                                      |
| 14 | 1sR: CAGCCGATTGGACGTAGGT                                                                                                       |
| 15 | 2sF: ACTCCTAAGTAAATGTGCA                                                                                                       |
| 27 | 2sR: CGAGGAGACGGTGACCAGGG                                                                                                      |

下面通过实施例，并结合附图，对本发明的技术方案作进一步详细的说明，但本发明不限于下面的实施例。

### 具体实施例

#### 5 实施例 1. 设计具有突变核苷酸和限制性核酸内切酶的酶切位点的寡聚物池

1.1 提供需要构建文库的氨基酸序列并确定其中需要突变的氨基酸的位置和种类

需要构建文库的氨基酸序列如下：RPDVNASX1X2X3GX4TPLHLAAX5X6GHLE  
10 IVEVLLKX7GADVVRPD (SEQ ID NO: 1)，其中 X1-X7 为需要突变的氨基酸，且 X1 突变为 20 种氨基酸中的 4 种，X2-X6 突变为 20 种氨基酸中的 16 种，X7 突变为 20 种氨基酸中的 3 种，因此，该文库库容为

4x16x16x16x16x16x3=12582912。所述氨基酸序列的编码核苷酸序列如下：  
AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTGGCTGCTX5X6GGTCACCT  
GGAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC (SEQ ID NO: 2),  
X1-X7 各代表 1 个待突变的密码子(3 个核苷酸)。

## 5 1.2 将序列分段

以单粘性末端为例，详述将序列分段的方法。

用单粘性末端将编码核苷酸序列分为两段。根据突变核苷酸的位置基于如下标准选择粘性末端的序列：1)粘性末端处于寡聚物序列中间，能将寡聚物池序列分为长度大致相同两段分别合成；2)粘性末端的序列中无回文序列；3)粘性末端序列的 GC 含量为 50%~75%；4)粘性末端的长度与选择的 IIS 型限制性核酸内切酶的酶切位点相匹配。

对于 1.1 的编码核苷酸序列，基于上述标准，由于 1) CCTG 位于 120 个核苷酸的 48-51 个；2) CCTG 中无回文结构；3) CCTG 的 GC 含量为 75%；4) CCTG 为 4 个核苷酸，对应于 IIS 型限制性核酸内切酶 *BsaI* 的酶切位点，  
15 因此选择 CCTG (在上述序列中用粗体表示)作为将序列分为两段的粘性末端。

## 1.3 分段设计各寡聚物池

第一段寡聚物长 51 bp，具体序列如下：5'-AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTG-3'(SEQ ID NO: 3)，第一段 DNA 序列的 3'端加入 *BsaI* 的识别序列 GGTCTC 的反向互补序列 GAGACC (SEQ ID NO: 4)，根据  
20 *BsaI* 特定的酶切位点，为了使 *BsaI* 酶切后产生选定的粘性末端 CCTG，在 *BsaI* 的识别序列的反向互补序列 GAGACC 的 5'端补加碱基 A，引入酶切位点后的序列为 AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTG**AGAGACC** (SEQ ID NO: 5)。在上述序列的 3'端进一步引入 18-21bp 的特异性序列以便后续扩增，所述特异性序列例如 ACCTACGTCCAAATCGGCTG (SEQ ID NO: 6)。第一  
25 段寡聚物的最终序列为 AGACCTGACGTTAACGCTAGCX1X2X3GGTX4ACTCCGCTGCACCTG**AGAGACC**ACCTACGTCCAAATCGGCTG (SEQ ID NO: 7)，命名为寡聚物池-1。

第二段寡聚物长 69 bp，具体序列如下：5'-GCTGCTX5X6GGTCACCTGGAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC-3' (SEQ ID NO: 8)，  
在第二段 DNA 序列的 5'端依次引入粘性末端序列 CCTG 和限制性核酸内切酶 *BsaI* 的识别序列 GGTCTC (SEQ ID NO: 9)，根据 *BsaI* 特定的酶切位点，  
30 为了使 *BsaI* 酶切后产生选定的粘性末端 CCTG，在 *BsaI* 的识别序列 GGTCTC

的 3'端后补加碱基 G, 得到的序列为 GGTCTC~~ACCT~~GGCTGCTX5X6GGTCACCTG  
GAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC (SEQ ID NO: 10)。

在上述序列的 5'端进一步引入 18-21bp 的特异性序列以便后续扩增, 所述  
特异性序列例如 *ACTCCTAAGTAAAAATGTGCAC* (SEQ ID NO: 11)。第二段寡聚物

- 5 的最终序列为 *ACTCCTAAGTAAAAATGTGCAC*GGTCTC~~ACCT~~GGCTGCTX5X6GGTCACCTG  
GAAATCGTTGAAGTTCTGCTGAAGX7GGTGCTGACGTGAGACCTGAC (SEQ ID NO: 12),  
命名为寡聚物池-2。

## 实施例 2. 通过 Library generate 程序设计寡聚物池

- 10 以单粘性末端为例, 一般地通过 Library generate 程序设计寡聚物池。

Library generate 程序网址为 [http://10.1.1.25/cgi-bin/aa\\_lib\\_generate.py](http://10.1.1.25/cgi-bin/aa_lib_generate.py),  
用于设计寡聚物池的 Library generate 程序的界面如图 1 所示。

该程序每次运行能够设计 1 个寡聚物池, 寡聚物池-1 和寡聚物池-2 需  
要分别设计。寡聚物池-1 的例示性设计过程如下:

- 15 在 Amino acid sequence 输入框中输入需要构建文库的氨基酸序列, 其中  
突变位点分别用数字<1>、<2>、<3>等表示, 一个数字代表一个氨基酸。对于  
寡聚物池-1, 输入的氨基酸序列为 RPDVNAS<1><2><3>G<4>TPLHLRDHLRPNRL;

- 在 DNA sequence 输入框中输入所述氨基酸序列的编码 DNA 序列, 突  
变位点分别用数字<1>、<2>、<3>等表示, 一个数字代表一个密码子。对  
20 于寡聚物池-1, 输入的 DNA 序列为 AGACCTGACGTTAACGCTAGC<1><2><3>GGT  
<4>ACTCCGCTGCACCTGAGAGACCACCTACGTCCAAATCGGCTG;

在 Size of the library 输入框中输入文库的序列数量。对于寡聚物池-1,  
<1>突变为 20 种氨基酸中的 4 种, <2>-<4>突变为 20 种氨基酸中的 16 种,  
寡聚物池-1 需要合成的寡聚物数量为  $4 \times 16 \times 16 \times 16 = 16384$  条, 因此输入 16384;

- 25 在 Expressing system 输入框中选择所需的表达体系, 如果在大肠杆菌  
中表达则选择“*E. coli*”, 如果在人细胞中表达则选择“Human”, 等等。对于  
寡聚物池-1, 采用人的密码子, 选择“Human”;

- 在“Upload the expected occurrence rates in excel file”选项中可以任选地上  
传预先存储的 excel 文件, 在该文件中可以设定寡聚物池-1 中每个突变的氨  
30 基酸的出现频率。

类似地, 寡聚物池-2 的例示性设计过程如下:

在 Amino acid sequence 输入框中输入需要构建文库的氨基酸序列，其中突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个氨基酸。对于寡聚物池-2，输入的氨基酸序列为 LLSKMCRSLAA<1><2>GHLEIVEVLLK<3>GADVRPD。

- 5 在 DNA sequence 输入框中输入所述氨基酸序列的编码 DNA 序列，突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个密码子。对于寡聚物池-2，输入的 DNA 序列为 CTCCTAAGTAAAATGTGCAGGTCTCGCCTGGCTGCT<1><2>GGTCACCTGGAAATCGTTGAAGTTCTGCTGAAG<3>GGTGTGACGTGAGACCTGAC。

- 10 在 Size of the library 输入框中输入文库的序列数量。对于寡聚物池-2，<1>-<2>突变为 20 种氨基酸中的 16 种，<3>突变为 20 种氨基酸中的 3 种，寡聚物池-2 需要合成的寡聚物数量为  $16 \times 16 \times 3 = 768$  条，因此输入 768；

在 Expressing system 输入框中选择所需的表达体系，对于寡聚物池-2，采用人的密码子，选择“Human”；

- 15 “在“Upload the expected occurrence rates in excel file”选项中可以任选地上上传预先存储的 Excel 文件，在该文件中可以设定寡聚物池-2 中每个突变的氨基酸的出现频率(期望值)。

表 1. 寡聚物池-1 和寡聚池-2 中每个突变氨基酸出现频率的设定

|        | X1  | X2 | X3 | X4 | X5 | X6 | X7  |
|--------|-----|----|----|----|----|----|-----|
| 丙氨酸 A  |     | 7% | 7% | 7% | 7% | 7% |     |
| 半胱氨酸 C |     |    |    |    |    |    |     |
| 天冬氨酸 D | 55% | 7% | 7% | 7% | 7% | 7% |     |
| 谷氨酸 E  |     | 7% | 7% | 7% | 7% | 7% |     |
| 苯丙氨酸 F |     | 4% | 4% | 4% | 4% | 4% |     |
| 甘氨酸 G  |     |    |    |    |    |    |     |
| 组氨酸 H  |     | 7% | 7% | 7% | 7% | 7% | 33% |
| 异亮氨酸 I |     | 4% | 4% | 4% | 4% | 4% |     |
| 赖氨酸 K  |     | 7% | 7% | 7% | 7% | 7% |     |
| 亮氨酸 L  |     | 4% | 4% | 4% | 4% | 4% |     |
| 甲硫氨酸 M |     |    |    |    |    |    |     |
| 天冬酰胺 N | 15% | 7% | 7% | 7% | 7% | 7% | 33% |
| 脯氨酸 P  |     |    |    |    |    |    |     |
| 谷氨酰胺 Q |     | 7% | 7% | 7% | 7% | 7% |     |
| 精氨酸 R  |     | 7% | 7% | 7% | 7% | 7% |     |
| 丝氨酸 S  | 15% | 7% | 7% | 7% | 7% | 7% |     |
| 苏氨酸 T  | 15% | 7% | 7% | 7% | 7% | 7% |     |
| 缬氨酸 V  |     | 4% | 4% | 4% | 4% | 4% |     |

|       |  |     |     |     |     |     |     |
|-------|--|-----|-----|-----|-----|-----|-----|
| 色氨酸 W |  | 4%  | 4%  | 4%  | 4%  | 4%  |     |
| 酪氨酸 Y |  | 10% | 10% | 10% | 10% | 10% | 33% |

### 实施例 3. 合成寡聚物池

通过本领域的已知的各种合成方法合成实施例 1 或实施例 2 设计的寡聚物池 1 和寡聚物池-2。所述方法包括但不限于磷酸三酯法、亚磷酰胺法、  
5 氢磷酸法以及通过芯片合成等等。

本发明中采用半导体芯片(CustomArray 公司的 B3P 合成仪)合成寡聚物池 1 和寡聚物池 2，分成两段序列合成寡聚物池只需要合成一张容量为 92k (92000 条序列)的芯片；如果不分成两段序列，合成一条长序列的寡聚物池的库容量达到  $1.26 \times 10^7$  需要多张芯片，芯片价格昂贵，因此，分段构建寡  
10 聚引物池大幅降低了构建成本。

### 实施例 4. 使用高保真 DNA 聚合酶对合成的寡聚物池进行扩增

根据上面得到的寡聚物池-1 和寡聚物池-2 的核苷酸序列，通过本领域已知的方法设计寡聚物池-1 和寡聚物池-2 的扩增引物，寡聚物池-1 的正向  
15 引物命名为 1F，反向引物命名为 1R。寡聚物池-2 的正向引物命名为 2F，反向引物命名为 2R。各引物的序列如下：

1F: AGACCTGACGTTAACGCTAG (SEQ ID NO: 13)

1R: CAGCCGATTTGGACGTAGGT (SEQ ID NO: 14)

2F: ACTCCTAAGTAAAATGTGCA (SEQ ID NO: 15)

20 2R: GTCAGGTCTCACGTCAGCAC (SEQ ID NO: 16)

PCR 反应体系配置如下：以合成的寡聚物池为模板，以 1F/1R 或者 2F/2R 为引物，加入高保真 DNA 聚合酶(例如 Phusion DNA 聚合酶)、dNTP、5x HF 缓冲液和牛血清白蛋白(BSA)，以水补齐 50  $\mu$ l，对合成的寡聚物池进行扩增，以引物 1F/1R 得到扩增后的寡聚物池-1，以引物 2F/2R 得到扩增后  
25 的寡聚物池-2。

反应体系如下：

|                |                     |
|----------------|---------------------|
| HF 缓冲液 (5x)    | 10 $\mu$ l          |
| 10 nM dNTP     | 1 $\mu$ l           |
| 模板             | 100 ng              |
| 正向/反向引物        | 2 $\mu$ l/2 $\mu$ l |
| BSA (20 mg/ml) | 10 $\mu$ l          |

|                        |                |
|------------------------|----------------|
| Phusion (2 U/ $\mu$ l) | 0.5 $\mu$ l    |
| H <sub>2</sub> O       | 添加至 50 $\mu$ l |

反应程序如下:

|   |                |
|---|----------------|
| 1 | 98 °C 30 秒     |
| 2 | 98 °C 10 秒     |
| 3 | 64 °C 10 秒     |
| 4 | 72 °C 20 秒     |
| 5 | 返回步骤 2, 16 个循环 |
| 6 | 72 °C 5 分钟     |

### 实施例 5. 使用 Golden Gate 将两个寡聚物池进行组装

- 配置 Golden Gate 组装体系, 加入寡聚物池-1 和寡聚物池-2, 在反应体系 5 系中同时加入 IIS 型限制性核酸内切酶 *Bsa*I 和 T4 DNA 连接酶两种酶, 以及 10x T4 DNA 连接酶缓冲液, 以水补齐至 20  $\mu$ l, 从而对两个寡聚物池进行组装。组装后的产物通过安捷伦 2100 检测, 检测结果如图 3 所示。

反应体系如下:

|                  |                |
|------------------|----------------|
| 寡聚物池-1           | 50 ng          |
| 寡聚物池-2           | 50 ng          |
| T4 DNA 连接酶缓冲液    | 2 $\mu$ l      |
| T4 DNA 连接酶       | 1 $\mu$ l      |
| <i>Bsa</i> I     | 1 $\mu$ l      |
| H <sub>2</sub> O | 添加至 20 $\mu$ l |

- 37°C 是限制性核酸内切酶 *Bsa*I 的最适反应温度, 16°C 是 T4 DNA 连接酶的最适反应温度, 在反应程序中, 先在 37°C 反应 3 分钟, 限制性核酸内切酶 *Bsa*I 识别并切割寡聚物池-1 和寡聚物池-2 序列中的 *Bsa*I 酶切位点, 各留下 4 bp 的粘性末端, 然后在 16°C 反应 5 分钟, T4 DNA 连接酶将两个寡聚物池在粘性末端处进行连接。连接正确的片段(即, 寡聚物池-1 与寡聚物池-2 相连的片段)上不含 *Bsa*I 识别序列, 连接错误的片段(即, 寡聚物池-1 与寡聚物池-1 或寡聚物池-2 与寡聚物池-2 相连的片段)含有 *Bsa*I 识别序列, 当再次在 37°C 反应时, 限制性核酸内切酶 *Bsa*I 能再次将其切开, 直至片段连接正确。进行 20 个循环后在 80°C 反应 15 分钟, 使 *Bsa*I 和 T4 DNA 连接酶失活, 以免对后续反应造成影响。最终得到组装后的寡聚物池。

反应程序如下:

|   |            |
|---|------------|
| 1 | 37 °C 3 分钟 |
| 2 | 16 °C 5 分钟 |

|   |                |
|---|----------------|
| 3 | 返回步骤 1, 20 个循环 |
| 4 | 80 °C 15 分钟    |
| 5 | 4 °C ∞         |

**实施例 6. 对组装后的寡聚物池进行扩增**

组装后的寡聚物池的序列如 SEQ ID NO: 17 所示。然而，组装后体系中的 DNA 量很低，纯化后基本得不到 DNA，且可能含有未组装的片段，  
5 因此需对组装后的序列进行 PCR 扩增。配置 PCR 反应体系如下：以组装后的寡聚物池为模板，以 1F/2R 为正向和反向引物，加入高保真 DNA 聚合酶（例如 Phusion DNA 聚合酶）、dNTP、5x HF 缓冲液，以水补齐至 50 μl，对组装后的寡聚物池进行扩增。

反应体系如下：

|                  |           |
|------------------|-----------|
| HF 缓冲液 (5x)      | 10 μl     |
| 10 nM dNTP       | 1 μl      |
| 模板(组装产物)         | 2.5ul     |
| 引物               | 1 μl/1 μl |
| Phusion (2 U/μl) | 0.5 μl    |
| H <sub>2</sub> O | 添加至 50 μl |

10 反应程序如下：

|   |                |
|---|----------------|
| 1 | 98 °C 30 秒     |
| 2 | 98 °C 10 秒     |
| 3 | 60 °C 30 秒     |
| 4 | 72 °C 20 秒     |
| 5 | 返回步骤 2, 25 个循环 |
| 6 | 72 °C 5 分钟     |

最终得到基因突变文库产物。

**实施例 7. 对基因突变文库产物进行纯化和测序**

对实施例 6 得到的产物通过琼脂糖凝胶电泳进行回收和纯化，得到了  
15 最终文库产物。最终文库产物的电泳图如图 4 所示。

将单粘性末端组装得到的最终文库产物平末端连接至经线性化的质粒 Puc57-EV，并使用通用引物 M13F 对终产物进行 Sanger 测序，检测序列是否正确；并进行高通量测序(NGS)，检测序列分布和氨基酸分布等。测序结果如图 5-图 7 和下表 1 所示。

图 5 是单粘性末端组装得到的文库序列的 Sanger 测序结果，该结果证明获得了所需的基因突变文库。

图 6 是单粘性末端组装得到的文库序列的 NGS 测序的序列分布图，表 1 是单粘性末端组装得到的文库序列的 NGS 测序的序列分布表。

5 表 2. 单粘性末端组装得到的文库序列的 NGS 测序的序列分布表

| 序列被检测到的次数 | 占总序列数量的百分比 | 计数      |
|-----------|------------|---------|
| 1         | 60.962325  | 4331739 |
| 2         | 22.300805  | 1584606 |
| 3         | 9.059291   | 643717  |
| 4         | 3.974710   | 282427  |
| 5         | 1.852300   | 131617  |
| 6         | 0.894773   | 63579   |
| 7         | 0.448182   | 31846   |
| 8         | 0.229959   | 16340   |
| 9         | 0.122565   | 8709    |
| 10        | 0.066891   | 4753    |
| 11        | 0.037506   | 2665    |
| 12        | 0.020927   | 1487    |
| 13        | 0.012328   | 876     |
| 14        | 0.007206   | 512     |
| 15        | 0.004447   | 316     |
| 16        | 0.002505   | 178     |
| 17        | 0.001393   | 99      |
| 18        | 0.000802   | 57      |
| 19        | 0.000394   | 28      |
| 20        | 0.000310   | 22      |
| 21        | 0.000169   | 12      |
| 22        | 0.000113   | 8       |
| 23        | 0.000056   | 4       |
| 24        | 0.000028   | 2       |
| 25        | 0.000014   | 1       |

图 7 是单粘性末端组装得到的文库序列的 NGS 测序的氨基酸分布情况。从上述数据可以看出，文库序列的氨基酸突变的位置、种类和频率的分布完全符合预期的设计。

10 实施例 8. 采用简并粘性末端组装获得基因突变文库

8.1 提供需要构建文库的氨基酸序列并确定其中需要突变的氨基酸的位置和种类

需要构建文库的氨基酸序列如下：KTEDTAVYYCSRDI02103104105106107  
DAWGQGLTVTVSS (SEQ ID NO: 18)，其中 AA102-AA107 为需要突变的氨基酸，  
5 且 AA102-AA106 突变为 20 种氨基酸中的 18 种，AA107 突变为 20 种氨基酸中的 4 种，因此，该文库库容为  $18 \times 18 \times 18 \times 18 \times 18 \times 4 = 7558272$ 。所述氨基酸序列的编码核苷酸序列如下：AAAACCGAGGACACGGCCGTGTATTACTGTAGTAG  
AGAT102103104105106107GACGCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCG  
(SEQ ID NO: 19)，AA102-AA107 各代表 1 个待突变的密码子(3 个核苷酸)。

## 10 8.2 分段设计各寡聚物池

将编码核苷酸序列分为两段，简并粘性末端选择为 AA104 的最后一个碱基和 AA105 的三个碱基。

第一段寡聚物长 51 bp，具体序列如下：5'-AAAACCGAGGACACGGCCGTGTAT  
TACTGTAGTAGAGAT102103104105-3' (SEQ ID NO: 20)，第一段 DNA 序列的 3'  
15 端加入 *BsaI* 的识别序列 GGTCTC 的反向互补序列 GAGACC (SEQ ID NO: 4)，  
为了精确切割，避免不需要序列的引入，在 *BsaI* 的识别序列的反向互补序列 GAGACC 的 5'端补加碱基 A，引入酶切位点后的序列为 AAAACCGAGGACACG  
GCCGTGTATTACTGTAGTAGAGAT102103104105AGAGACC (SEQ ID NO: 21)。在上述序列的 3'端进一步引入 18-21bp 的特异性序列以便后续扩增，所述特异性  
20 性序列例如 *ACCTACGTCCAAATCGGCTG* (SEQ ID NO: 6)。第一段寡聚物的最终  
序列为 AAAACCGAGGACACGGCCGTGTATTACTGTAGTAGAGAT102103104105AGAGACC  
*ACCTACGTCCAAATCGGCTG* (SEQ ID NO: 22)，命名为寡聚物池-1。

由于 Library generate 设计序列需要输入氨基酸序列和 DNA 序列，所以设计第二段寡聚物时需要先加上 AA104 的前两个碱基，因此第二段寡聚物  
25 长 51 bp，具体序列如下：5'-104105106107GACGCCTGGGGCCAAGGAACCCTGGTC  
ACCGTCTCCTCG-3' (SEQ ID NO: 23)，在设计完序列之后，再去掉 AA104 的前两个碱基。由于第二段寡核苷酸的特殊性(需要去掉 AA104 的前两个碱基)，因此，要先使用 Library generate 设计出所有序列，再添加 *BsaI* 酶切位点/特异性序列等 DNA 序列。

## 30 8.3. 通过 Library generate 程序设计寡聚物池

Library generate 程序网址为 [http://10.1.1.25/cgi-bin/aa\\_lib\\_generate.py](http://10.1.1.25/cgi-bin/aa_lib_generate.py),

用于设计寡聚物池的 Library generate 程序的界面如图 1 所示。

该程序每次运行能够设计 1 个寡聚物池，寡聚物池-1 和寡聚物池-2 需要分别设计。寡聚物池-1 的例示性设计过程如下：

5 在 Amino acid sequence 输入框中输入需要构建文库的氨基酸序列，其中突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个氨基酸。对于寡聚物池-1，输入的氨基酸序列为 KTEDTAVYYCSR<1><2><3><4>RDH LRPNRL；

在 DNA sequence 输入框中输入所述氨基酸序列的编码 DNA 序列，突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个密码子。对于寡聚物池-1，输入的 DNA 序列为 AAAACCGAGGACACGGCCGTGTATTACTGTAGT  
10 AGAGAT<1><2><3><4>AGAGACCACCTACGTCCAAATCGGCTG；

在 Size of the library 输入框中输入文库的序列数量。对于寡聚物池-1，<1>-<4>突变为 20 种氨基酸中的 18 种，寡聚物池-1 需要合成的寡聚物数量为  $18 \times 18 \times 18 \times 18 = 104976$  条，因此输入 104976；

15 在 Expressing system 输入框中选择所需的表达体系，如果在大肠杆菌中表达则选择“*E. coli*”，如果在人细胞中表达则选择“Human”，等等。对于寡聚物池-1，采用人的密码子，选择“Human”；

在“Upload the expected occurrence rates in excel file”选项中可以任选地上上传预先存储的 excel 文件，在该文件中可以设定寡聚物池-1 中每个突变的  
20 氨基酸的出现频率。

序列 1 经过 *Bsa*I 酶切后产生的粘性末端位于 AA104 的最后一个碱基和 AA105 的三个碱基，为简并粘性末端，具体为 CAAC/CAAG/CACC/CAGC/CATC/CCAC/CCAG/CCCC/CCGG/CCTG/CGAC/CGAG/CGCC/CGGC/CGTG/CTAC/CTGG/CTTC/GAAC/GAAG/GACC/GAGC/GATC/GCAC/GCAG/GCCC/GCGG/GCTG/GGAC/GGAG/GGCC/GGGC/GGTG/GTAC/GTGG/GTTC，共 36 种。  
25

类似地，寡聚物池-2 的例示性设计过程如下：

在 Amino acid sequence 输入框中输入需要构建文库的氨基酸序列，其中突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个氨基酸。对于寡聚物池-2，输入的氨基酸序列为<1><2><3><4>DAWGQGLVTVSS。

30 在 DNA sequence 输入框中输入所述氨基酸序列的编码 DNA 序列，突变位点分别用数字<1>、<2>、<3>等表示，一个数字代表一个密码子。对

于寡聚物池-2，输入的 DNA 序列为<1><2><3><4>GACGCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCG。

在 Size of the library 输入框中输入文库的序列数量。对于寡聚物池-2，<1>-<3>突变为 20 种氨基酸中的 18 种，<4>突变为 20 种氨基酸中的 4 种，寡聚物池-2 需要合成的寡聚物数量为  $18 \times 18 \times 18 \times 4 = 23328$  条，因此输入 23328；

在 Expressing system 输入框中选择所需的表达体系，对于寡聚物池-2，采用人的密码子，选择“Human”；

“在“Upload the expected occurrence rates in excel file”选项中可以任选地上上传预先存储的 excel 文件，在该文件中可以设定寡聚物池-2 中每个突变的氨基酸的出现频率(期望值)。

在软件输出序列之后，使用 excel 处理得到的数据，使用 excel 中的分列功能将 AA104 的前两个碱基和之后的序列分开，删除 AA104 的前两个碱基记为 104'。之后在第二段 DNA 序列的 5'端依次引入粘性末端序列 CCTG 和限制性核酸内切酶 *BsaI* 的识别序列 GGTCTC (SEQ ID NO: 9)，为了精确切割，避免不需要序列的引入，在 *BsaI* 的识别序列 GGTCTC 的 3'端后补加碱基 G，得到的序列为 GGTCTCG104'105106107GACGCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCG (SEQ ID NO: 24)。在上述序列的 5'端进一步引入 18-21bp 的特异性序列以便后续扩增，所述特异性序列例如 *ACTCCTAAGTAAAATGTGCAC* (SEQ ID NO: 11)。第二段寡聚物的最终序列为 *ACTCCTAAGTAAAATGTGCAC*GGTCTCG104'105106107GACGCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCG (SEQ ID NO: 25)，命名为寡聚物池-2。寡聚物池-2 在 *BsaI* 酶切后产生的粘性末端和寡聚物池-1 产生的粘性末端相同，具体为 CAAC/CAAG/CACC/CAGC/CATC/CCAC/CCAG/CCCC/CCGG/CCTG/CGAC/CGAG/CGCC/CGGC/CGTG/CTAC/CTGG/CTTC/GAAC/GAAG/GACC/GAGC/GATC/GCAC/GCAG/GCCC/GCGG/GCTG/GGAC/GGAG/GGCC/GGGC/GGTG/GTAC/GTGG/GTTC，共 36 种。寡聚物池-1 与寡聚物池-2 通过相同的粘性末端连接。

表 3. 寡聚物池-1 和寡聚池-2 中每个突变氨基酸出现频率的设定

|        | AA102 | AA103 | AA104 | AA105 | AA106 | AA107 |
|--------|-------|-------|-------|-------|-------|-------|
| 丙氨酸 A  | 7%    | 7%    | 7%    | 7%    | 7%    |       |
| 半胱氨酸 C |       |       |       |       |       |       |
| 天冬氨酸 D | 6%    | 6%    | 6%    | 6%    | 6%    |       |

|        |     |     |     |     |     |     |
|--------|-----|-----|-----|-----|-----|-----|
| 谷氨酸 E  | 3%  | 3%  | 3%  | 3%  | 3%  |     |
| 苯丙氨酸 F | 3%  | 3%  | 3%  | 3%  | 3%  | 25% |
| 甘氨酸 G  | 17% | 17% | 17% | 17% | 17% |     |
| 组氨酸 H  | 2%  | 2%  | 2%  | 2%  | 2%  |     |
| 异亮氨酸 I | 3%  | 3%  | 3%  | 3%  | 3%  | 25% |
| 赖氨酸 K  | 1%  | 1%  | 1%  | 1%  | 1%  |     |
| 亮氨酸 L  | 5%  | 5%  | 5%  | 5%  | 5%  | 25% |
| 甲硫氨酸 M |     |     |     |     |     | 25% |
| 天冬酰胺 N | 3%  | 3%  | 3%  | 3%  | 3%  |     |
| 脯氨酸 P  | 5%  | 5%  | 5%  | 5%  | 5%  |     |
| 谷氨酰胺 Q | 2%  | 2%  | 2%  | 2%  | 2%  |     |
| 精氨酸 R  | 7%  | 7%  | 7%  | 7%  | 7%  |     |
| 丝氨酸 S  | 11% | 11% | 11% | 11% | 11% |     |
| 苏氨酸 T  | 6%  | 6%  | 6%  | 6%  | 6%  |     |
| 缬氨酸 V  | 5%  | 5%  | 5%  | 5%  | 5%  |     |
| 色氨酸 W  | 4%  | 4%  | 4%  | 4%  | 4%  |     |
| 酪氨酸 Y  | 10% | 10% | 10% | 10% | 10% |     |

#### 8.4 简并粘性末端基因文库的构建

按照本发明实施例 3 的方法合成寡聚物池。根据上面得到的寡聚物池-1 和寡聚物池-2 的核苷酸序列，通过本领域已知的方法设计寡聚物池-1 和寡聚物池-2 的扩增引物，寡聚物池-1 的正向和反向引物分别命名为 1sF 和 1sR。

5 寡聚物池-2 的正向和反向引物分别命名为 2sF 和 2sR。各引物序列如下：

1sF: AAAACCGAGGACACGGCCGT (SEQ ID NO: 26)

1sR: CAGCCGATTTGGACGTAGGT (SEQ ID NO: 14)

2sF: ACTCCTAAGTAAAAATGTGCA (SEQ ID NO: 15)

2sR: CGAGGAGACGGTGACCAGGG (SEQ ID NO: 27)

10 以合成的寡聚物池为模板，以 1sF/1sR 或者 2sF/2sR 为引物，加入高保真 DNA 聚合酶(例如 Phusion DNA 聚合酶)，按照本发明实施例 4 的方法对合成的寡聚物池进行扩增，以引物 1sF/1sR 得到扩增后的寡聚物池-1，以引物 2sF/2sR 得到扩增后的寡聚物池-2。

15 按照本发明实施例 5 的方法使用 Golden Gates 方式将两个寡聚物池进行组装。按照本发明实施例 6 的方法对组装后的寡聚物池进行扩增，最终文库产物的电泳图如图 8 所示。按照本发明实施例 7 的方法对基因文库产物进行纯化和测序。测序结果如图 9-图 11 和下表 2 所示。

图 9 是简并粘性末端组装得到的文库序列的 Sanger 测序结果, 该结果证明获得了所需的基因突变文库。

图 10 是简并粘性末端组装得到的文库序列的 NGS 测序的序列分布图, 表 2 是简并粘性末端组装得到的文库序列的 NGS 测序的序列分布表。

5 表 4. 简并粘性末端组装得到的文库序列的 NGS 测序的序列分布表。

| 序列被检测到的次数 | 占总序列数量的百分比  | 计数     |
|-----------|-------------|--------|
| 1         | 34.72731754 | 620006 |
| 2         | 13.83322308 | 246972 |
| 3         | 8.600282761 | 153546 |
| 4         | 6.152197881 | 109839 |
| 5         | 4.710222188 | 84094  |
| 6         | 3.759990661 | 67129  |
| 7         | 3.080016084 | 54989  |
| 8         | 2.578414205 | 46034  |
| 9         | 2.189449654 | 39089  |
| 10        | 1.88074763  | 33578  |
| 11        | 1.631243758 | 29123  |
| 12        | 1.425476997 | 25450  |
| 13        | 1.258155311 | 22463  |
| 14        | 1.114544016 | 19899  |
| 15        | 0.996043945 | 17783  |
| 16        | 0.891604083 | 15918  |
| 17        | 0.800186778 | 14286  |
| 18        | 0.721688263 | 12885  |
| 19        | 0.656627366 | 11723  |
| 20        | 0.598414984 | 10684  |
| 21        | 0.544820162 | 9727   |
| 22        | 0.498073855 | 8892   |
| 23        | 0.457605354 | 8170   |
| 24        | 0.420768642 | 7512   |
| 25        | 0.388082545 | 6929   |
| 26        | 0.356901047 | 6372   |
| 27        | 0.330337108 | 5898   |
| 28        | 0.308183198 | 5502   |
| 29        | 0.286963176 | 5123   |
| 30        | 0.264134791 | 4716   |
| 31        | 0.245197608 | 4378   |

|     |             |       |
|-----|-------------|-------|
| 32  | 0.229736566 | 4102  |
| 33  | 0.215053763 | 3839  |
| 34  | 0.200319079 | 3576  |
| 35  | 0.188282294 | 3362  |
| 36  | 0.175363503 | 3131  |
| 37  | 0.162963539 | 2909  |
| 38  | 0.152690766 | 2726  |
| 39  | 0.143196233 | 2557  |
| 40  | 0.135569477 | 2420  |
| >41 | 2.689910109 | 48024 |

图 11 是简并粘性末端组装得到的文库序列的 NGS 测序的氨基酸分布情况。从上述数据可以看出，文库序列的氨基酸突变的位置、种类和频率的分布完全符合预期的设计。

5 还需要说明的是，在可实施且不明显违背本发明的主旨的前提下，在本说明书中作为某一技术方案的构成部分所描述的任一技术特征或技术特征的组合同样也可以适用于其它技术方案；并且，在可实施且不明显违背本发明的主旨的前提下，作为不同技术方案的构成部分所描述的技术特征之间也可以以任意方式进行组合，来构成其它技术方案。本发明也包含在  
10 上述情况下通过组合而得到的技术方案，并且这些技术方案相当于记载在本说明书中。

以上通过具体实施方式和实施例对本发明进行了说明，但本领域技术人员应该理解的是，这些并非意图对本发明的范围进行限定，本发明的范围应由权利要求书确定。

15

### 工业实用性

本发明提供了一种大库容文库的构建方法，能通过合成较少的寡聚物序列通过组装变成大库容的文库，大大的节约了大库容的文库构建的成本，解决了合成大库容文库价格昂贵的问题。

20