摘 要

近几十年来，随着测序技术和分子标记技术的发展，生命科学研究取得了长足的进步，特别是在作物育种领域，二代测序技术和分子标记技术更是为遗传机理的解析、新品种的培育和品种改良提供了巨大的助力。而普通小麦（*Triticum aestivum* L.）作为全球超过 35%人口的主粮，其育种研究更是重中之重。但是，普通小麦作为异源六倍体，其基因组异常复杂，导致了小麦遗传育种研究远远落后于基因组较小的水稻和玉米。随着 2018 年小麦第一版高质量参考基因组序列和基因组注释信息的公布，小麦功能基因组研究和育种研究进入了新时代。无论是育种研究还是功能基因组研究，全基因组基因型检测（Whole genome genotyping, WGG）都是一个非常重要的研究工具，特别是在遗传图谱构建和 QTL 定位中，WGG 更是不可缺的。虽然目前市场上已有多种基于 SNP 芯片和二代测序技术的 WGG 被应用于小麦的育种研究中，但这些方法成本太高，限制了其广泛应用，尤其是在成百上千个样品的育种项目中，WGG 的成本更是很多育种工作者不能接受的。因此昂贵的 WGG 已经成为了当前小麦育种研究的一大阻碍。为了解决这一难题，我们主要进行了以下研究，并获得了如下的结果：

（1）优化了反向限制性位点相关 DNA 测序技术（inverse Restriction-site Associated DNA sequencing，iRAD-seq），并将优化后的 iRAD-seq 测序技术与全基因组低覆盖度测序技术结合起来，开发出了低覆盖度反向简化基因组测序技术（Skimmed inverse Restriction-site Associated DNA sequencing，siRAD-seq）。该技术的核心思想是利用 iRAD-seq 和低覆盖度测序两种手段对小麦基因组进行两次简化，主要实验流程包括：第一，通过模拟酶切，选出合适的限制性内切酶；第二，对要进行 WGG 的样本构建基于 Tn5 转座酶的重测文库；第三，使用 All-in-one sequencing（AIO-seq）的混库技术，对所有样本混库；第四，酶切混库；第五，分选，低覆盖度测序。该技术相较于传统的简化基因组测序技术主要有以下优点：第一，采用了“先建库后酶切的模式”，节约了酶的用量，也简化了实验步骤；第二，使用了 AIO-seq 混库技术，使各样本的测序数据有较好的均一性，同时也节省了混库成本；第三，使用了低覆盖度测序技术，降低了测序和数据分析的成本；第四，大量的使用了国产试剂和仪器，将 WGG 的成本降低至￥50~100/样。

（2）从临麦 2 号×中麦 892 小麦 RIL 群体中随机抽取 5 个子代，分别使用 10×WGS，55 K 芯片和不同数据量的 siRAD-seq 进行 WGG，比较 10×WGS 与其他方式 WGG 结果的一致性，并将这种一致性定义为 WGG 的准确性。分析结果发现，siRAD-seq 在 WGG 中的准确性远高于 55 K 芯片，并且 siRAD-seq 的测序数据量（数据量：0.5 ~5 Gb，准确性：0.925~0.975）对 WGG 准确性的影响不大。无论是哪种方式，B 亚基因组 WGG 的准确性高于亚基因组 A 和亚基因组 D。这可能与小麦亚基因组的复杂度有关。

（3）分别使用小麦 90 K SNP 芯片和 siRAD-seq 对两个小麦 RIL 群体进行 WGG 并根据 WGG 的结果，分别构建遗传图谱，并定位 5 个与产量相关性状的 QTL。WGG 的结果发现，siRAD-seq 在两个小麦 RIL 群体中分别可以检测到 4,152,066 和 3,748,768 个高质量的 SNP 位点（双亲间同时检测到高质量的 SNP 位点），虽然大多数子代 SNP 的缺失率高于 30%，但所有子代的 SNP 都均匀的分布在基因组上，即使子代测序数据量仅为 1 Gb，在群体 SNP 缺失率为 20%时，依旧可以检测到~280 K 个 SNP 位点。在构建遗传图谱时，siRAD-seq 相较于 90 K 芯片可以获得更多的标记，并且每一个标记在参考基因组上都有唯一且确定的物理位置，所以获得的遗传图谱质量更高。在定位 QTL 时，基于 90 K 芯片定位到的 QTL 的物理区间的均值是基于 siRAD-seq 的 2-4 倍。并且我们也发现，不同测序数据量（0.5~5 Gb）的 siRAD-seq 对临麦 2 号×中麦 892 群体 WGG 的影响不大。

（4）使用 siRAD-seq 研究了 10 个普通小麦-中间偃麦草属间杂交的衍生后代染色体的重构。研究结果发现，siRAD-seq 可以检测到衍生后代中中间燕麦草染色体片段的导入，同时，我们还检测到这些衍生后代中小麦染色体的缺失/加倍。同样的材料使用 FISH 检测时，仅能发现部分中间偃麦草染色体片段的导入和小麦染色体的缺失/加倍，很难发现部分染色体片段的缺失/加倍，不能发现染色单体的缺失/加倍。相较于 FISH，siRAD-seq 检测的成本更低，实验流程更简单，准确性更高。

关键词：siRAD-seq，全基因组基因型检测，普通小麦，QTL 定位，小麦-中间偃麦草属间杂交的衍生后代

目 录

缩写词（Abbreviations）

英文缩写	英文全称	中文名称
AIO-seq	All-In-One sequencing	高通量混库测序技术
BLUE	Best linear unbiased estimation	最佳线性无偏估计
CS	Chinse Spring	中国春（小麦品种）
gDNA	Genomic DNA	基因组 DNA
KL	Kernel length	粒长
KNS	Kernel number per spike	穗粒数
KW	Kernel width	粒宽
LOD	Logarithm of odds	概率对数
MAB	marker-assisted breeding	标记辅助育种
R^2	Phenotypic variance explained	表型解释力
RIL	Recombinant inbred line	重组自交系
SN	Spike number per unit area	单位面积穗数
SNP	Single Nucleotide Polymorphism	单核苷酸多态性
TKW	Thousand-kernel weight	千粒重
WGG	Whole genome genotyping	全基因组基因型检测

1 前 言

近几十年来，随着测序技术、分子标记技术、生物信息学和大数据技术的发展，作物育种已经逐渐的从传统育种（2G 和 3G 育种）过渡到了分子育种（4G 育种）。在可预期的未来，智能育种技术（5G 育种）将成时代的主流，但在当前的技术条件下，4G 育种技术才是最有希望被大规模应用到作物育种中的技术[1]。而要将 4G 技术育种推广到作物育种中，无论如何也绕不开全基因组基因型检测（Whole genome genotyping, WGG）。目前依赖于 SNP 芯片和二代测序技术的 WGG，在一些小基因组作物育种中得到了广泛的应用，但像小麦（17 Gb）这种基因组超大的作物[2]，依赖于 SNP 芯片和二代测序技术 WGG 的成本是绝大多育种工作者不能承受的。所以开发并优化一款高效，低成本的小麦 WGG 技术对小麦育种是十分重要的。而随着测序技术的快速发展，测序成本以不可思议的速度降低，只有将测序技术充分的应用到小麦的 WGG 中，才是将来的发展之道。

1.1 测序技术的发展

1953 年，Watson 和 Crick 在 Nature 杂志上发表了 DNA 的双螺旋结构[3]，标志着生命科学研究进入了分子生物学时代，人类对生命也有了新的见解，即：“生命是序列的，生命是数字的”[4]。因此可以破解生命序列为目标的测序技术便应用而生，从 1977 年出现的以 Sanger 测序法为代表的第一代测序技术到今天以纳米孔测序技术为代表的第三代测序技术，测序技术在不断的革新，并为生命科学和生产力的发展做出了不可磨灭的贡献。

1.1.1 第一代测序技术

20 世纪 70 年代，由 Sanger 和 Coulson 提出的链终止法测序技术和由 Maxam 和 Gilbert 提出的链降解法（化学法）测序技术开创了测序的先河[5, 6]，为后续测序技术的发展鉴定了坚实的基础。其中链终止法测序技术也被称为 Sanger 测序法，是一种依靠带有特殊标记的双脱氧核苷酸（ddNTPs）终止 DNA 链合成而获得 DNA 序列信息的测序技术。其核心原理为：当反应体系中有四种单脱氧核苷酸（dNTPs）、引物、待测 DNA 片段和 DNA 聚合酶时，在适宜的反应条件下能进行正常的聚合反应，但将 dNTPs

换成 ddNTPs 时，因为 ddNTPs 缺乏延伸时所需要的 3-OH 基团，DNA 延伸时不能形成完整的磷酸二酯键，导致 DNA 合成终止。所以 Sanger 测序法便依据此原理在 4 个反应体系中分别加入一定比例的含有特殊标记的 ddNTPs 进行聚合反应，当遇到 ddNTP 时反应会终止，会形成起始位置相同但终止位置各异、长度不一的 DNA 片段，并且终止位置均为带有特殊标记的 ddNTP，然后通过电泳区分出不同长度的 DNA 片段，再根据每条 DNA 片段末端的碱基便可确定待测 DNA 片段的序列信息。在开发后的不久，这种测序技术就被应用到科学研究中，例如，1977 年 Sanger 等人使用这种测序方法获得了第一个物种的（噬菌体 X174，基因组大小：5,375 bp）基因组序列[7]。随着技术的进步，Sanger 测序法也在不断的发展，1986 年，Smith 等人使用非放射性的荧光标记代替同位素标记，可通过分析电泳过程中激发出的荧光信号实现了 Sanger 测序法的自动化操作[8]，而毛细管电泳技术的发明更是大大的提高了测序的速度。

化学降解法测序技术是一种分别采用不同的化学试剂对特定碱基进行化学修饰并在该位置打断核苷酸链而获得 DNA 序列信息的技术。其核心原理为：首先对待测 DNA 片段的一端进行放射性标记，然后在不同的反应体系中分别加入不同的化学修饰试剂（不同的试剂作用不同的碱基）进行碱基断裂反应，形成一系列起点相同但长度不一的 DNA 片段，最后通过电泳和放射自显影技术获得待测 DNA 片段的序列信息。但随着测序技术的发展，化学降解法测序技术因其技术的局限性和高的成本退出了历史的舞台。

自 1977 年第一代测序技术问世到如今，Sanger 测序技术为生命科学的发展做出了不可磨灭的贡献，例如，第一个真核生物（酿酒酵母）[9]、第一个古细菌（詹氏甲烷球菌）[10]、第一个多细胞生物（大肠杆菌）[11]和人的基因组序列[12]都是使用 Sanger 测序法解密的。今天，虽然 Sanger 测序技术的测序速度和通量已经远远落后于二代和三代测序技术，但因其较长的读长和极高的准确度依旧活跃在克隆测序、单基因测序和癌症研究等领域中。

1.1.2 第二代测序技术

随着多个物种的基因组序列被解密，生命科学研究已经进入了后基因组学时代，导致测序的需求不断增加，但第一代测序技术因其低的通量和高的成本已经越来越不能满足测序的需求。面对这一窘况，第二代测序技术（Next-generation Sequencing，

NGS）应运而生，其最显著的特点是通量高、价格低。在众多二代测序技术中主要有 4 中测序技术：Roche 公司的焦磷酸测序技术、ABI 公司的 SOLiD（Supported Oligo Ligation Detection）测序技术、Illumina（Solexa）公司的 SBS 测序技术，和 BGI 公司的 DNBSEQ™测序技术，随着技术的迭代，前两种测序技术已经被淘汰，后两种测序技术占据了绝大多数的市场份额。

2005 年，发表在 Nature 上的焦磷酸测序技术开创了二代测序（边合成边测序）的先河[13]。焦磷酸测序技术的核心原理为：当待测的 DNA 片段在 DNA 聚合酶的作用下，加入的碱基可以和模板匹配并释放出一个焦磷酸分子，焦磷酸在 ATP 磷酸化酶的催化下与底物（APS）合成 ATP，ATP 在荧光素酶的作用下释放出不同的信号，将这些信号采集和解码，便可获得待测 DNA 片段的序列信息。基于此原理，454 Life Sciences 公司于同年年底率先推出了第一款二代测序平台：Genome Sequencer 20 System，接着该公司推出了更好的测序平台：Genome Sequencer FLX System，2008 年 Roche 公司在前两款测序平台的基础上开发出通量更高、读长更长、准确度更高的测序平台：GS FLX Titanium。虽然焦磷酸测序技术因其较长的读长（400-500 bp）曾一度占据了测序界的半壁江山，但随着 Illumina 公司推出了通量更高、成本更低的测序平台，在 2014 年，焦磷酸测序技术随着 Roche 公司退出了测序领域。

2007 年，ABI 公司推出了 SOLiD 测序技术，该测序技术的核心原理为：酶连接法测序（Sequencing by Oligo Ligation Detection）。具体过程如下：测序引物（n）、连接酶和分别使用 4 种颜色的荧光染料（CY5、Texas Red、CY3，6-FAM）标记的 8 碱基单链荧光探针（结构为：3’-XXnnnzzz-5’，其中 XX 是确定的碱基，nnn 是随机碱基，zzz 是可以与任何碱基配对的特殊碱基，并且荧光标记就加在此处）和模板链反应时，当探针与模板匹配而被连接上时，代表 XX 碱基的荧光信号会被激发，记录荧光信号后，再通过切割探针的 5-6 位的磷酸二酯键来淬灭荧光，以便进行下一个位置的测序。以此方式测序时，每一次都与上一次有 5 个碱基的 Gap，即第一次测序的位置为 1-2，第二次为 6-7，以此类推直到测序结束，然后进行第二轮测序，第二轮的测序引物为 n-1，所以测序的位置为 2-3，按照此种方式，进行 5 轮测序，待测 DNA 将被完全测序，并且每个位置都能被测序两次（这也是 SOLiD 测序准确度高的重要原因）[14]。SOLiD 测序文库的长度仅为焦磷酸测序的四分之一，虽然 SOLiD 测序技术的通量远高于焦磷酸测序（Emulsion PCR 时，微滴远远小于焦磷酸测序），准确度也远高于同期的其他测

序平台，但 SOLiD 测序技术因读长较短（仅为焦磷酸测序技术的四分之一）、价格昂贵、技术迭代缓慢、测序流程复杂于 2012 年退出了测序领域。

2006 年，Illumina 公司通过收购 Solexa 公司进入了测序领域，并在随后的几十年间不断的推出了各种通量的测序平台，其中最具代表性的测序平台有：Mi-seq 系列、Hi-Seq 系列和 NovaSeq 系列[15]。无论哪种测序平台，它的核心原理都是边合成边测序（Sequencing by Synthesis，SBS）技术，测序的过程大致可以分为 4 个步骤：第一步是构建待测样品的 DNA 文库，通常使用机械打断或转座酶将待测 DNA 打断成片段长度为 200-800 bp 的小片端，接着在打断后的 DNA 片段两端添加上测序接头并经过 PCR 扩增形成待测文库；第二步是将构建好的 DNA 文库变性成单链并随机的附着在 Flow cell（流动槽）上；第三步是通过桥式 PCR 将固定好的 DNA 文库扩增成簇，起到放大信号的作用；第四步是测序，主要依靠两种技术，即 SBS 测序技术和 3’端可逆屏蔽终止子技术（dNTPs 的 3’-OH 位置是被修饰过的，在测序时每次只能由一个碱基与模板结合，测序完成后经过处理这个化学修饰基团可被去掉，以便进行下一轮测序）。所以在测序时，当新合成的链在模板、引物、酶和经过修饰的 dNTPs 的作用下延伸时，每次只能结合一个核苷酸，接着洗脱未结合的核苷酸，并激发结合的核苷酸带的信号，并拍照记录信号，然后淬灭信号并通过 3’端可逆屏蔽终止子技术去掉 3’端的修饰基团，如此循环直到测序完成，最后通过计算机处理所获得的信号，便可获得待测 DNA 文库的序列信息[16]。在过去的十几年里，Illumina 测序平台凭借着超高的通量、高准确度和低成本一直占据着很高的市场份额，时至今日，虽然三代测序技术已经取得了长足的进步，但 Illumina 测序依旧在变异检测、RNAseq、单细胞测序、产前筛查和肿瘤检测等领域占据着主要的市场份额。

2013 年，华大基因也通过收购 Complete Genomics 公司进入了测序领域。并在随后的十年间不断创新，推出了一系列具有自主知识产权的高通量测序平台，特别是 2023 年 2 月推出的 DNBSEQ-T20×2 测序平台更是刷新了测序通量和成本的纪录，将单个人全基因测序成本降低至 100 美元以下。相较于其他测序技术，华大测序技术最大的区别便是自研的 DNBSEQ™测序技术，其测序过程可以分为以下几个步骤：第一步是构建测序文库（类似于 Illumina 文库）；第二步是将构建好的双链文库通过高温变性等方式转化为单链状态的环状文库；第三步是以环状文库为模板通过滚环复制制作成纳米球（DNA nanoballs，DNBs），并将 DNBs 加载到测序芯片上；第四步借助联合探针锚定