天津中医药大学本科毕业论文

绞股蓝皂苷生物合成的候选糖基转移酶筛选研究

The Screening of Glycosyltransferases Involved in Gypenoside Biosynthesis

学　　院　中药学院

专　　业　中药资源与开发

学生姓名　孙思杰

学　　号　201414022015

指导教师　王丽芝

天津中医药大学

2018 年 05 月

天津中医药大学学位论文原创性与诚信声明

本人郑重声明：所呈交的学位论文《　　》是本人在导师指导下独立进行的研究工作和取得的研究成果。除了文中特别加以标注引用和致谢之处外，论文中不包含其他个人或集体已经发表或撰写的研究成果。也不包含获得天津中医药大学或其他教育机构的学位、学历使用过的材料。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示了谢意。

本人郑重承诺：与本论文相关的研究数据及成果，其知识产权归属指导教师所属单位。未经指导教师及其所属单位同意，本人不得擅自以任何形式发表相关论文及成果，不得擅自从事与课题有关的任何开发和盈利活动。

本人完全清楚本声明的法律后果，申请学位论文和资料若有不实之处，本人愿承担相应的法律责任。

学位论文作者签名：　　日期：　年　月　日

学位论文版权使用授权书

本学位论文作者完全了解天津中医药大学有关保留、使用学位论文的规定，特授权天津中医药大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校保留并向国家有关部门或机构送交论文的复印件和电子版。

本论文属于

1、□保密，在____年解密后适用本授权书。

2、□不保密。

请在以上相应方框内打“ √ ”

学位论文作者签名：　　导师签名：

日期：　年　月　日　　日期：　年　月　日

目录

摘要

目的：绞股蓝皂苷是绞股蓝的主要活性成分，本研究对绞股蓝中参与绞股蓝皂苷生物合成的尿苷二磷酸糖基转移酶(UGT)基因进行挖掘，以期为后续更多药用植物无参全长转录组研究提供新思路。**方法：**测序平台选用二代与三代相结合，将二者数据混合之后去冗余。挑选注释到 UGT 的序列，提取相应二代数据定量结果，使用进化树亲缘关系分析、热图聚类分析筛选。选出可能性最高的糖基转移酶候选序列进行荧光定量 PCR 验证。**结果：**1) 测序数据共 98.32Gb，非冗余转录本共 140,157 条，注释到 UGT 的序列有 254 条；2) 结合进化树亲缘关系与热图分析分析，挑选出 6 条进行荧光定量 PCR 验证；3) 根据荧光定量 PCR 结果，推测出 1 条最有可能的候选基因。**结论：**1) 本研究建立了一种基于无参全长转录组二三代混合测序技术，挑选次生代谢产物合成相关基因的分析流程；2) 基于此分析流程，我们认为 GpUGT35 为最有可能参与绞股蓝皂苷生物合成的候选基因。

关键词： 绞股蓝；绞股蓝皂苷；全长转录组；尿苷二磷酸糖基转移酶

ABSTRACT

Objective: Gypenosides, a group of triterpenoid saponins, are the main active ingredient from *Gynostemma pentaphyllum*. In this study, we first combined PacBio isoform sequencing and Illumina sequencing (Hybrid-sequencing) to mine the UGTs (UDP-Glycosyltransferase) involved in the biosynthesis and regulation of gypenosides. May the pipeline we developed provide a possible solution for other non-reference Iso-Seq research of medical plants. **Methods:** Illumina NextSeq 500 and PacBio RSII platform were employed to generate raw data for analysis. After removing the redundancies, expression analysis results were used to generate heatmaps, and phylogenetic analysis was combined to screen the candidate UGTs for real-time PCR. **Results:** Using 98.32 Gb transcriptome data from three different tissues, roots, stems leaves, a total of 140,157 unigenes were generated with an average length of 750bp. To further identify the genes involved in the synthesis and regulation of gypenosides, we used these unigenes search against databases and obtained 254 UGTs. After the phylogenetic analysis, tissue-specific expression and the expression response to MeJA-treated in leaves, we speculated that GpUGT35 was one of the candidate genes for the biosynthesis of gypenosides. **Conclusion:** GpUGT35 might involve the biosynthesis of gypenosides. We further described a new procedure to mine UGT genes involved in the gypenoside biosynthesis by hybrid sequencing of *G. pentaphyllum* transcriptome.

Keywords: *Gynostemma pentaphyllum*; Gypenosides; Isoform Sequencing; UDP-Glycosyltransferase

论文正文

1 绪言

1.1 绞股蓝皂苷

绞股蓝(*Gynostemma pentaphyllum*)是一种分布于亚洲的多年生草质攀援植物，主要生长在中国、日本、韩国印度、缅甸、孟加拉、印度尼西亚、马来西亚的等地，生长地区海拔高度约为60m-3200m[1]。绞股蓝皂苷中是其主要活性成分之一。目前已明确结构的绞股蓝皂苷有 201 种，可按苷元相似程度分为 13 类[2]。绞股蓝皂苷的苷元部分均为达玛烷型四环三萜类，其中，原人参二醇的含量最高。糖基主要连接于 C-3、C-6 和 C-20 位，类型主要有葡萄糖、阿拉伯糖、鼠李糖、木糖等，依据连接糖的个数可分为单糖、二糖和三糖[3]。

近年来对于人参有效成分抗肿瘤活性的研究表明，苷元的活性要高于皂苷。绞股蓝皂苷在水解后的苷元部分有许多与人参苷元活性一致的组份，提示其有着重大的药用价值。在对于绞股蓝皂苷及其水解产物结构的研究中，更多的是运用酸水解的方式，除此之外还有碱水解、酶水解以及生物转化的方法；在对其药理活性的研究中，研究方向多集中在抗肿瘤活性方面。对于如何利用绞股蓝水解产物的结构优势、发现更多先导化合物，正在受到国内外学者越来越多的关注[4]。

1.2 植物糖基转移酶

糖基转移酶是指通过合成糖苷键，将糖基连接到特定受体的一类酶。该类酶广泛存在于各种原核生物、真核生物、古生物和病毒中，能够识别不同受体、供体并形成多种多样的产物。其中，与植物次生代谢相关的多为糖基底物经过尿苷二磷酸修饰的糖基转移酶(UDP-Glycosyltransferase, UGT)[5]。

UGT 在植物中的功能，首先是在植物次生代谢途径中完成次生代谢产物合成的最后一步加糖，不仅可以增加小分子苷元的可溶性，还可以稳定他们的构型。其次是调节植物体内激素水平，在激素活性维持、存储及运输方面有着重要作用。它也是一些活性分子不可缺少的组成部分，失去糖基会让该类分子彻底失去活性。

本研究通过对于测序后数据的深入分析，提供对于绞股蓝皂苷生物合成中候选糖基转移酶的预测，为绞股蓝皂苷的后续研究以及大规模工业化生产提供了相关的数据与理论支持。

1.3 全长转录组测序

转录组从广义上讲，是指细胞或组织内全部 RNA 的总和，而总 RNA 依据不同的分类标准有着不同的类型，如依据是否翻译成蛋白而分为编码 RNA (coding RNA)和非编码 RNA (non-coding RNA)；依据长短而分为长 RNA (long RNA, 长度大于 200bp)和短 RNA (small RNA, 长度小于 200bp)；依据翻译中行使不同的功能主要分为信使 RNA (message RNA, mRNA)、转运 RNA (transfer RNA, tRNA)、核糖体 RNA (ribosomal RNA, rRNA)。由于 mRNA 研究较多，一般的转录组测序主要针对于 mRNA[7]，与此同时也可以测到部分长非编码 RNA (long non-coding RNA, lncRNA)。

二代高通量测序技术在转录组方面的应用被称为 RNA 测序(RNA Sequencing, RNA-Seq)，RNA-Seq 由于其高通量的优势以及日益下降的成本，在转录组方面的应用到现在在研究领域中已经基本取代了基因芯片(micro array)技术。然而，二代测序首先读长较短，现在应用最多的是双端 150bp 测序，测出的片段需要拼接才可以形成转录本，对于转录本的还原度较差；其次，在建库中经过多次 PCR 扩增，容易造成差异分析中假阳性概率上升[8]。

三代测序技术尤其是 PacBio 公司基于单分子实时测序技术(Single Molecule Real Time sequencing, SMRT)为转录组打造的全长转录组测序(Isoform Sequencing, Iso-Seq)流程，已经成为国内外研究中新的热点。其在优势主要表现在读长较长，所测得数据无需组装从而可以减少拼接错误，但是价格昂贵又限制了其测序通量无法做到像二代数据一样高[9]。

本研究在分析中选用三代测序技术数据与二代数据相结合，将二代数据与三代数据混合之后去冗余，而非仅用二代数据进行矫正和定量分析，减少了一般分析中对于二代数据的浪费。为三代无参转录组测序数据分析提供了新方法。

2 材料与方法

2.1 材料、试剂与设备

2.1.1 测序材料与数据

测序所用的绞股蓝幼苗购买自江西萍乡，移栽至营养土中，24℃进行 16 小时光照和 8 小时暗培养。取经过 200μM 0.8% 乙醇配制的茉莉酸甲酯(MeJA)处理 0 小时，6 小时，12 小时，24 小时的绞股蓝叶，以及未处理过的根、茎、叶三个部位样品，立即放于液氮之中，-80℃保存备用[10]。

三代测序数据使用 PacBio RSⅡ测序平台获得；二代测序数据使用 Illumina NextSeq

500 测序平台获得。测序工作由北京诺禾致源生物信息科技有限公司完成。

2.1.2 荧光定量 PCR 试剂

1) 天根 RNAprep Pure 多糖多酚植物总 RNA 提取试剂盒(TIANGEN, DP441)；
2) GoScript™ Reverse Transcription System 试剂盒(Promega, A5001)；
3) SYBR® Premix Ex Taq™ (Takara, 6109)。

2.1.3 主要设备

1) 台式离心机(Thermo Fisher Scientific Legend Micro 17R)；
2) Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific)；
3) 微量移液枪(Eppendorf: 2.5μl, 10μl, 20μl, 100μl, 200μl, 1000μl)；
4) 实时荧光定量 PCR 仪(Bio-RAD CFX96 Real Time System)。

2.2 实验方法

2.2.1 测序结果处理

使用经过 SMRT Analysis software 处理的三代数据非冗余转录本和经过 Trinity[11, 12] 拼接后的非冗余转录本，使用 CD-HIT 软件[13]进行聚类去冗余，阈值设定在 85%。聚类去冗余后得到的非冗余转录本(Unigenes)进行下面的功能注释。

2.2.2 功能注释

对于所有经过 CD-HIT 的 Unigenes，功能注释主要使用 KEGG (http://www.kegg.jp，代谢通路注释), SwissProt (http://www.uniprot.org/uniprot，蛋白质功能注释)， 以及 Pfam (https://pfam.xfam.org，蛋白质结构域比对)，三个数据库进行注释。在糖基转移酶筛选时使用了 CAZy (http://www.cazy.org，碳水化合物相关酶数据库，糖基转移酶注释)与一部分上述三个数据库的注释结果，注释结果通过糖基转移酶相关关键词筛选，为下游分析缩小范围，降低分析难度。

2.2.3 注释结果筛选

取所有的 Unigenes 进行 Pfam 注释后注释到 UDP-GT 的序列号和糖基转移酶注释中(本研究中选用 dbCAN: http://cys.bios.niu.edu/dbCAN2/) 注释到 GT1 的序列号；将两部分注释结果取并集，取其中长度大于 300aa 的序列以去除过短的序列。结果用于提取二代数据定量结果、对应序列用于进化树亲缘关系分析。

2.2.4 亲缘关系分析

使用文献中已经进行过功能验证的人参皂苷合成酶的蛋白序列，与本研究中挑选到的糖基转移酶的蛋白序列使用 MEGA 6.0 软件[17]建立进化关系树，进行同源性分析。

综合进化树亲缘分析与聚类热图结果，挑选与已经鉴定的基因同源性较高或与上游基因表达模式相似的糖基转移酶序列，设计引物进行 qPCR 验证。

2.2.5 基于二代数据的表达量聚类分析

2.2.5.1二代数据定量

使用二代数据的短 reads 对 Unigenes 进行回帖定量，所用软件为 RSEM[14]。定量结果使用 edgeR 包[15]的 TMM 均一化方法进行不同处理与不同部位定量结果组间均一化，所得值用于聚类展示。

2.2.5.2热图聚类分析

在 NCBI 蛋白数据库中下载到绞股蓝皂苷合成的上游基因(FPS, SS, SE)的蛋白序列，对 Unigenes 翻译后的蛋白序列进行 BLAST 比对，提取出对应 Unigene ID，从而提取表达量定量结果。之后再挑选出所有注释到糖基转移酶的 Unigenes 表达量定量结果。

对本节上述所有的 Unigenes 表达量使用 pheatmap 包[16]进行聚类与热图绘制。在展示中，TMM 值经过对数处理，使用公式为 $\log_e(TMM+1)$。

2.2.6 cDNA 获取

2.2.6.1RNA 提取

参照天根 RNAprep Pure 多糖多酚植物总 RNA 提取试剂盒(DP441)操作说明，略有改动，具体操作步骤如下：

1) 取 475μl 裂解液 SL，加入 25μl β-巯基乙醇，混匀备用。
2) 取 50-100 mg 样品在液氮中迅速研磨成粉末。
3) 转移粉末至 2ml RNase-Free 离心管中并加入 1 中裂解液。
4) 12,000 rpm 离心 2 分钟。
5) 上清转移至过滤柱 CS 上，12,000 rpm 离心 2 分钟。
6) 上清转移至新的 RNase-Free 的离心管中，加入 0.5 倍体积的无水乙醇，混匀。
7) 转入吸附柱 CR3 中，12,000 rpm 离心 15 秒， 弃去废液后将吸附柱 CR3 放回收集管中 。
8) 向吸附柱 CR3 中加入 350μl 去蛋白液 RW1，12,000 rpm 离心 15 秒。
9) 向 CR3 加入 80μl DNase I 工作液，室温放置 15 分钟。
10) 重复 8。
11) 向 CR3 中加入 500μl 漂洗液 RW 12,000 rpm 离心 15 秒，弃去废液后 CR3 放回收集管中。

12) 重复 11。

13) 12,000 rpm 空离 2 分钟。

14) 将吸附柱 CR3 放入一个新的 RNase-Free 离心管中，悬空滴加 50μl RNase-Free ddH_2O，室温放置 2 分钟。

15) 12,000 rpm 离心 1 分钟，所得 RNA 样品取 1μl 用于琼脂糖凝胶电泳检测， 1μl 用于 Nanodrop 浓度测定。

2.2.6.2 反转录

参照 GoScript™ Reverse Transcription System 试剂盒操作说明，具体操作步骤如下：

1) 在 200μl 离心管中配制如下反映混合液：

反应体系组分	体积
RNA Samples	≤ 5μl
Oligo dT Primer (50μM)	1μl
dNTP Mixture (10mM each)	1μl
RNase-Free ddH_2O	up to 5μl

2) 70℃保温 5min 后，冰上迅速冷却至少 5min，离心 10 秒，置于冰上。

3) 依次配制下列反应液，总量 15μl。

反应体系组分	体积
变性后反应液	5μl
GoScript 5×Reaction Buffer	4μl
$MgCl_2$	3μl
PCR Nucleotide Mix	1μl
Recombinant RNasin Ribonuclease Inhibitor	0.5μl
GoScript Reverse Transcriptase	1μl
Nuclease-Free water	5.5μl

4) 使用移液枪混匀

5) PCR 反转，条件如下：

退火：25℃ 5min

延伸：42℃ 60min

变性：70℃ 5min

6) 反转成功的 cDNA 保存于-20℃，-80℃长期保存备份。

2.2.7 引物设计

对于持家基因、绞股蓝皂苷合成上游基因以及通过 2.2 的方法筛选到的糖基转移酶，设计引物如下：

基因名称	引物序列
ACTIN-F	CCGAGTGGCCCCTGAAGAG
ACTIN-R	AAGTATGGCATGGGGGAGAGC
HMGR-F	ACCAATGCCGTTTTCTTCAC
HMGR-R	ATCGACCGTTCATCGTCTTC
FPS-F	CTGGGTCTGCTTTCCCATAA
FPS-R	TTGTTATGGCGGGTGAAAAT
SS-F	ACAGCTTCAGCCTCAGCTTC
SS-R	CATGAAAAATGCCAGTCACG
SE-F	TGGCTTCCACCATAAACACA
SE-R	AACTTAACGGGCGAGGATTT
GpUGT1-F	ATAGGACCAAACGTGCCATC
GpUGT1-R	AAAGCTGCATAGCTCCCAAA
GpUGT8-F	TGTCTTGGAAACCATCACGA
GpUGT8-R	AGAGCTCAAAACCTCGTCCA
GpUGT16-F	CATCGGTGATTTACGTGTCG
GpUGT16-R	GCGAACACCACGGAACTATT
GpUGT22-F	TCCCACTCATCGAATTCTCC
GpUGT22-R	GATTTTGGAGCCTTGTGGAA
GpUGT35-F	GGAACCCTTTCGGTAATGCT
GpUGT35-R	GTTTTCGACGGTGTTCGTTT
GpUGT44-F	CAACACCCCTTCACTTTCGT
GpUGT44-R	TAGCCCGGGTAACTGTATGC

2.2.8 荧光定量 PCR

该实验使用 Takara SYBR® Premix Ex Taq™，于 Bio-RAD CFX96 Real Time System 完成，具体反应体系及步骤如下：

1) cDNA 模板稀释至 100ng，引物稀释至 10μM。

2) 冰上配制以下反应体系：

反应体系组分	体积
2×SYBR Premix Ex Taq	10μl
Forward primer (10μM)	0.5μl
Reverse primer (10μM)	0.5μl
cDNA	1μl
Nuclease-Free water	up to 20μl

3) 进行实时荧光定量 PCR 分析，条件如下：

95℃	30s	Stage 1
95℃	5s	Stage 2 × 40 Cycles
60℃	30s	
72℃	15s	
95℃	10s	Stage 3
4℃	∞	Stage 4

3 结果

3.1 基于二三代测序平台混合测序的绞股蓝转录组分析

3.1.1 测序结果

为了尽可能多的获取高质量的非冗余转录本，我们采取了不同于一般测序公司的分析策略，使用二代与三代数据混合去冗余后继续进行下游分析(图 3.1)。在三代数据中，测序共使用了 8 个 cell，下机数据大小为 12.50Gb。原始下机数据使用 PacBio 官方的 SMRT Analysis software 处理，得到 268,927 条环形一致性序列，这些序列根据是否具有完整的 5’非翻译区，Poly-A 位点以及 3’非翻译区而被进一步分类为 99,739 条全长非嵌合序列(分到全长序列的序列会在程序中进行筛选，如果一条中有两个 3’非翻译区则会被判定为嵌合序列而被过滤掉)和 142,079 条非全长序列。全长非嵌合序列会经过 ICE 的聚类而将不同 ZMW 孔中测到的同一条转录本只保留一条，之后这些经过孔间自校正后的序列会同时使用 Quiver 使用非全长的序列对 ICE 去冗余后的序列进行矫正从而得到高质量一致性序列与低质量一致性序列两个部分的序列，为了不浪费低质量(只是得到非全长序列的支持较少)的序列，我们将两部分序列进行聚类去冗余(CD-HIT)最终得到了 32,426 条经过校正转录本和 8,550 条非冗余转录本。

二代测序数据方面，经过茉莉酸甲酯诱导不同时间与根茎叶不同部位的测序原始数据总计有 85.82Gb，去除接头后和低质量 reads 之后的高质量数据总计 576,532,682 条。为了得到尽可能多的非冗余转录本，这些数据被共同送入 Trinity 进行拼接，得到了 140,601 条 Trinity 拼接出的非冗余转录本。

为了对二代和三代数据有一个直观的了解，我们对上述步骤中三代数据经过矫正的转录本 (32,426 条)、非冗余转录本(8,550 条)、Trinity 使用二代数据拼接出的非冗余转录本(140,601 条)进行了统计并绘制了小提琴图(图 3.2a)。如图所示，二代的数据量明显大于三代数据，而且大部分长度分布在 1000bp 以下；对于三代数据，数据量虽然较小，但是平均长度明显长于于二代数据，经过去冗余后的非冗余转录本平均长度还有了明显的提升。

由于 8,550 条三代非冗余转录本和 140,601 条 Trinity 拼接出的非冗余转录本之间还会有冗余，所以我们使用 CD-HIT 软件在二者之间去冗余，最终我们得到了 140,157 条非冗余转录本(Unigenes，下文中如非特别指出，提到非冗余转录本或 Unigenes 皆指向此处的 140,157 条 Unigenes)，以这些 Unigenes 进行之后的下游分析。

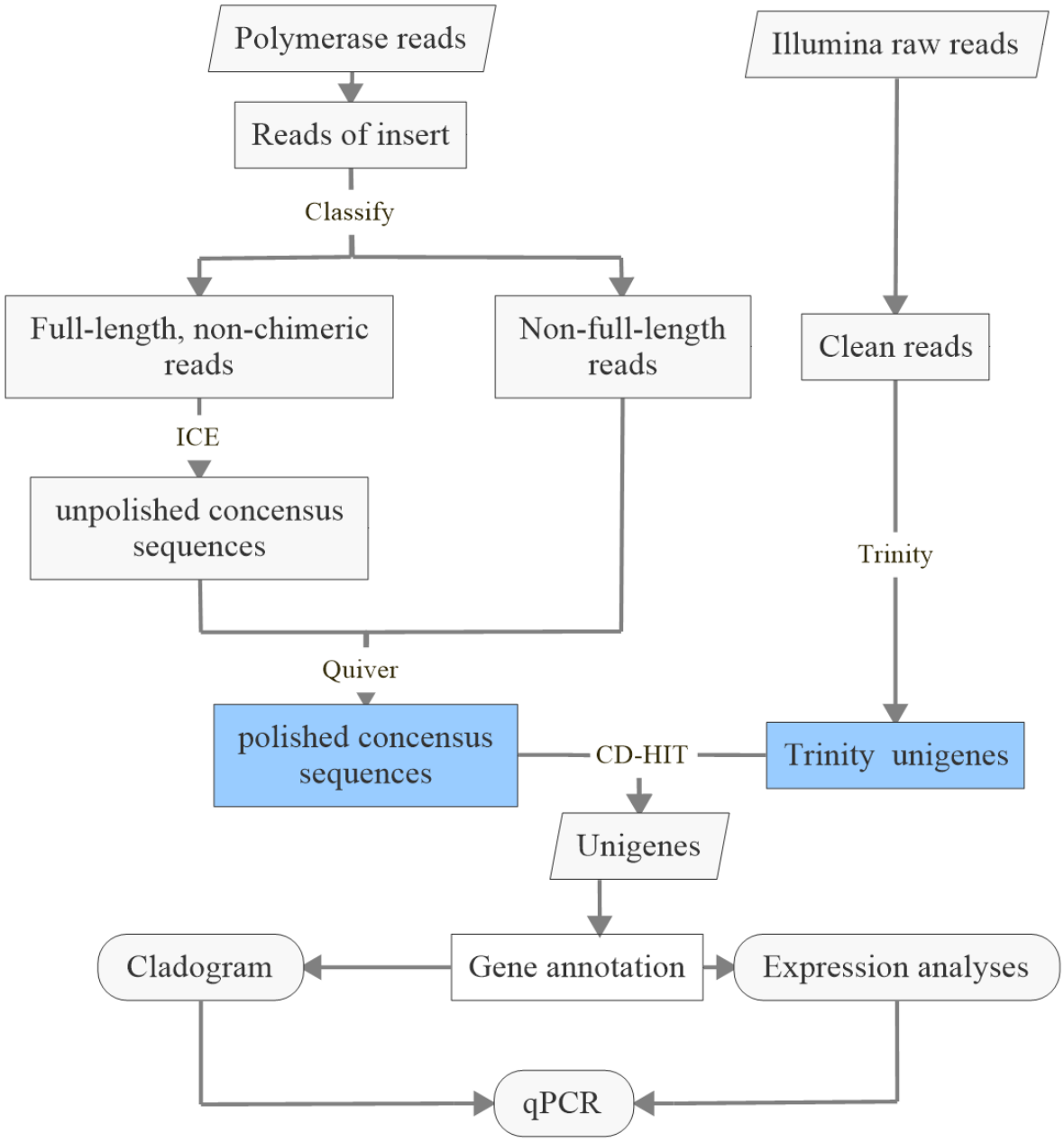

图 3.1 主要技术路线 平行四边形代表了主要的数据输入与输出，阴影部分代表了本研究分析中主要的数据来源，椭圆形代表了主要的筛选步骤，而箭头上标注了所进行的主要处理。

3.1.2 功能注释

对所有的 Unigenes 使用 SwissProt、KEGG 和 Pfam 三大数据库进行注释。从注释结果中可以看出，有 68,692 (49%)条的 Unigenes 至少在三个数据库之一中得到了注释，有 25,959 条序列在三个数据库之中均有注释(图 3.2b)。在 KEGG 数据库中，45,381 条序列被注释到五种代谢通路(细胞功能、环境信息处理、遗传信息处理、新陈代谢、人类疾病)中，我们挑选前四种与本研究有关的通路进行展示(图 3.2c)。

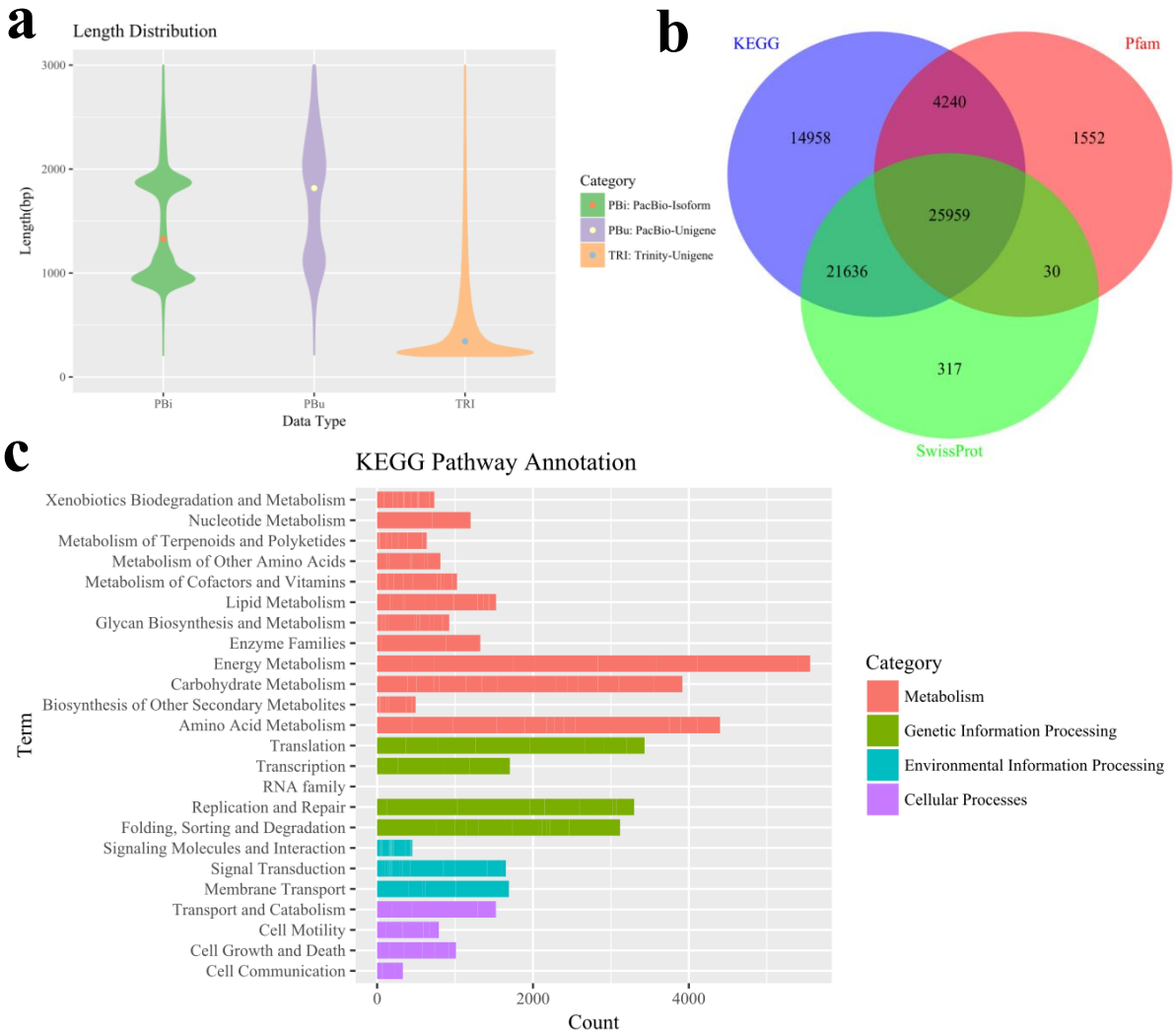

图 3.2 测序数据统计与注释　(a) 进行了三代数据校正后转录本、非冗余转录本、二代数据拼接转录本的长度分布统计，横坐标为转录本类型，而纵坐标为转录本长度(bp)，图形的宽度代表着在此长度分布的转录本数量。(b) 对所有经过 CD-HIT 的 Unigenes 进行注释，在三大数据库的注释情况如图所示。(c) KEGG 注释结果展示，挑选了四大分类下的二级条目做数量统计，横坐标为注释到纵坐标对应条目的 Unigenes 数量，纵坐标为注释到的二级条目。

其中，与能量代谢相关的 Unigenes 数量超过五千，其次是氨基酸代谢和碳水化合物代谢。此外，有 614 条 Unigenes 注释到了萜类代谢通路中，以及 190 条与萜类骨架生物合成相关。这些注释信息为后续候选基因的挖掘和绞股蓝皂苷的生物合成相关研究提供了数据支持。

3.2 进化树亲缘关系分析

68 条 GpUGT 可以依据序列相似性阈值来分到 20 个 UGT 家族中，分类标准为：相似性大于 40%为同一家族，相似性大于 60%为同一亚家族，其中有 9 条绞股蓝 UGT (*Gynostemma pentaphyllum* UGT, GpUGT)可以被分到新的 UGT 家族中(图 3.3)。

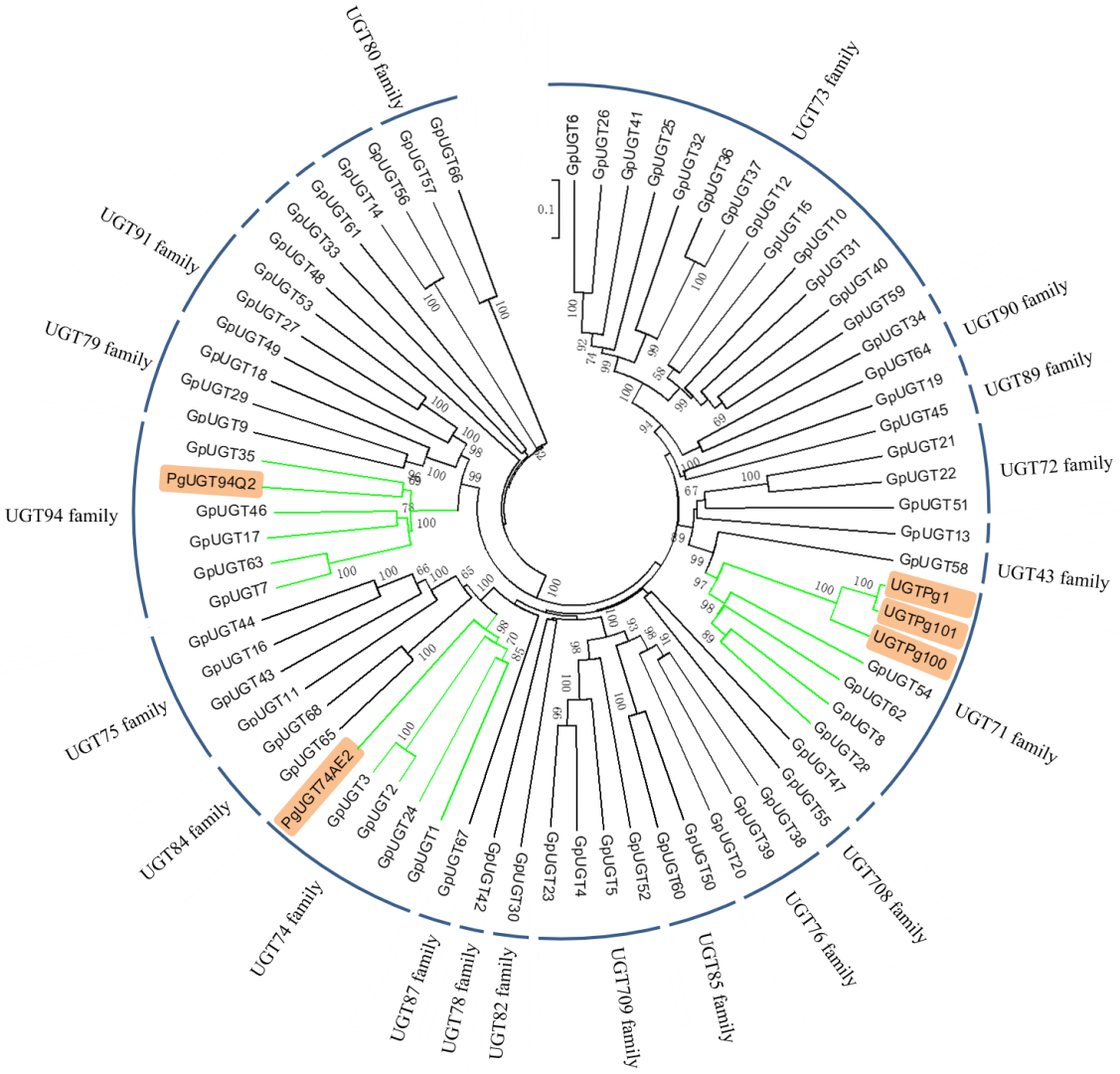

图 3.3 UGT 进化树　　将 GpUGT 按照序列相似性分为不同的家族，在最外环标出。阴影部分为人参中鉴定过的五条人参 UGT，与它们聚类到同一支上的 GpUGT 有着较高的序列相似性，说明有着更相近的功能。

进化树亲缘关系分析显示(图 3.3)，共有 13 条 GpUGT 与已经鉴定功能的五条人参 UGT (Panax ginseng UGT, PgUGT)聚类到了进化树的同一分支上，其中，GpUGT35 与 PgUGT94Q2 聚到同一支上，相似度最高，达到了 50%以上；GpUGT1 与 GpUGT24 和 PgUGT74AE2 在同一分支上；GpUGT8 与 UGTPg1 (UGT71 家族)聚到了一起，它们的相似性都在 40%以上。

在根据序列相似性构建的进化树上，相似性越高，越趋向于聚类在同一分支上，表明这些序列更有可能具有相似的功能。所以根据聚类结果，我们挑选了上述序列中的 GpUGT1、GpUGT8、GpUGT35 进行后续分析。

3.3 基于二代测序数据的表达量分析

UGT 是绞股蓝皂苷合成途径的最后一步关键酶，注释结果中注释到 UGT 的 Unigenes 共有 254 条，因为有的序列过短(<300aa)，故推测其并非全长的 UGT 序列，所以我们选择其中氨基酸序列长度在 352aa-524aa 的 68 条候选序列(GpUGT1-GpUGT68)进行下游分析。

经过二代数据定量之后，进一步提取 68 条 GpUGT 的表达量，进行热图聚类分析(图 3.4a, b)。分析结果显示，上游基因 SS、SE 在不同组织、不同部位的表达量均偏高，并且有部分候选 GpUGT 与上游基因聚类到同一支上。

为了更清晰的展示聚类结果，我们将与上游基因聚类到同一支上的 GpUGT 挑选出来，单独聚类展示(图 3.4c, d)并在图中标注了计算后的表达量值。不难看出，在不同诱导条件处理中，GpUGT4, GpUGT16, GpUGT35, GpUGT44 与上游基因聚到了同一支上，提示有可能受到 SS 与 SE 的调控(图 3.4c)。在不同组织部位中，SS 与 SE 在叶中的表达量明显高于根中和茎中，而 GpUGT18 和 GpUGT22 也符合这一趋势，故同样被列入侯选序列之中。

综合上述结果，我们挑选了与上游基因表达趋势最为相似的 GpUGT16、GpUGT22、GpUGT44 进行后续分析。

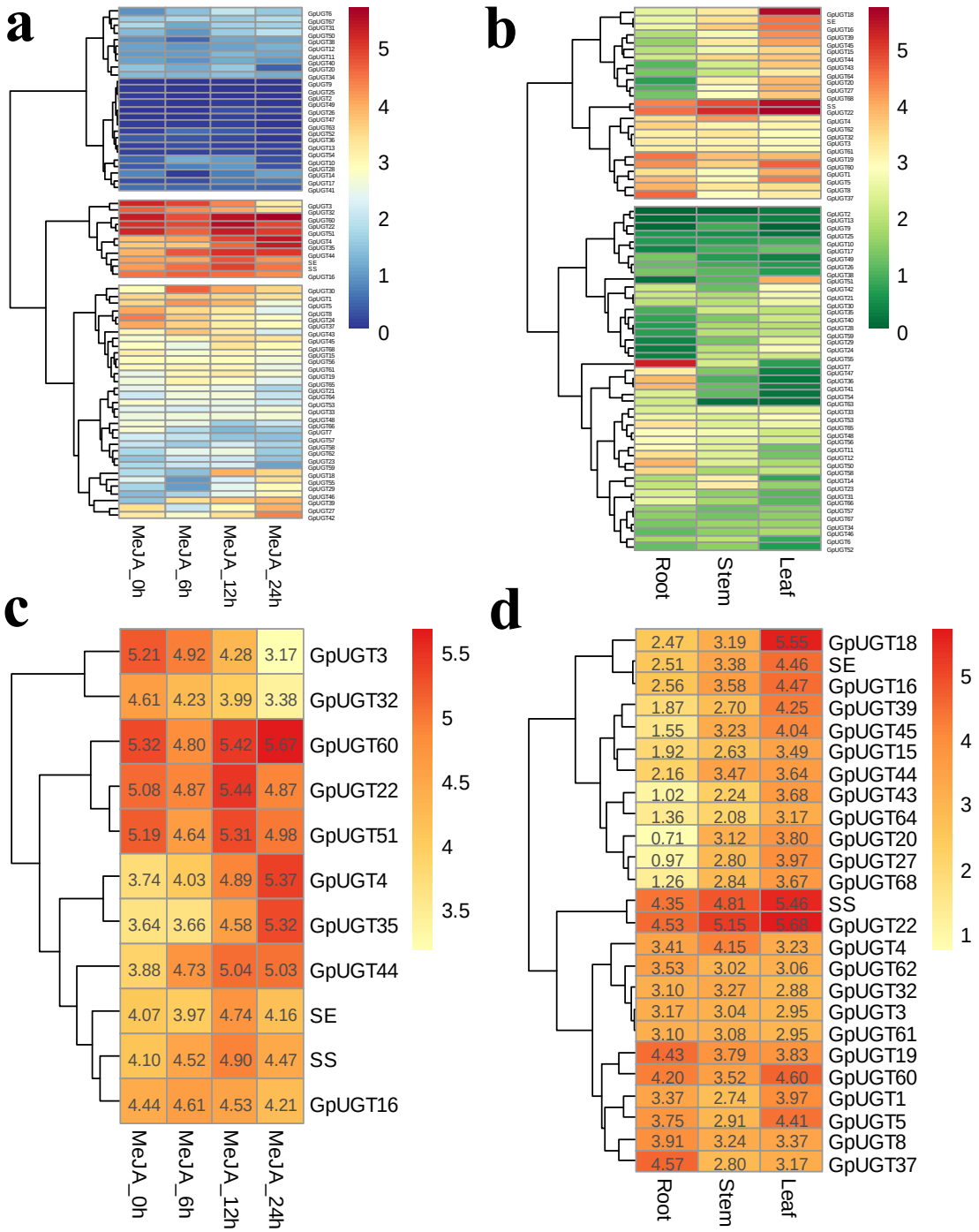

图 3.4 UGT 表达量聚类热图　(a) 茉莉酸甲酯诱导处理后不同时间的 GpUGT 表达量展示，其中，红色代表表达量较高，蓝色代表表达量较低。(b) GpUGT 在不同组织部位的表达量展示，其中，红色代表表达量较高，绿色代表表达量较低。(c) 诱导后不同时间后表达量与上游基因在 a 中聚到相同支的 GpUGT 再聚类结果。(d) 不同组织部位 GpUGT 在 b 中与上游基因聚到同一支上的再聚类展示。上述所有的 TMM 值都经过 log 处理，具体方法见 2.2.5.2。

3.4 基于荧光定量 PCR 的表达量分析

综合热图聚类与进化树聚类结果，我们挑选了六个候选 GpUGT 进行荧光定量 PCR 验证(GpUGT 1, GpUGT8, GpUGT16, GpUGT22, GpUGT35, GpUGT44)。并对其荧光定量 PCR 结果进行绘图展示。

在不同部位中(图 3.5a)，上游基因(HMGR, FPS, SS, SE)表达量较高并且聚类到了一起。其中，GpUGT35 与上游基因表达模式最为相似，但是表达量较低。

在不同诱导条件处理中(图 3.5b)，GpUGT1 和 GpUGT8 在 6 小时处受到了明显的诱导上调；GpUGT22 和 GpUGT35 在整体表达趋势上与上游基因较为相似，均为 6 小时处表达量不高甚至受到抑制，而在 12 至 24 小时处出现增长；除 GpUGT16 表达量持续走低之外，其余五个 GpUGT 均受到诱导而升高。特别值得注意的是，GpUGT35 在表达趋势上同样与上游基因最为相似。

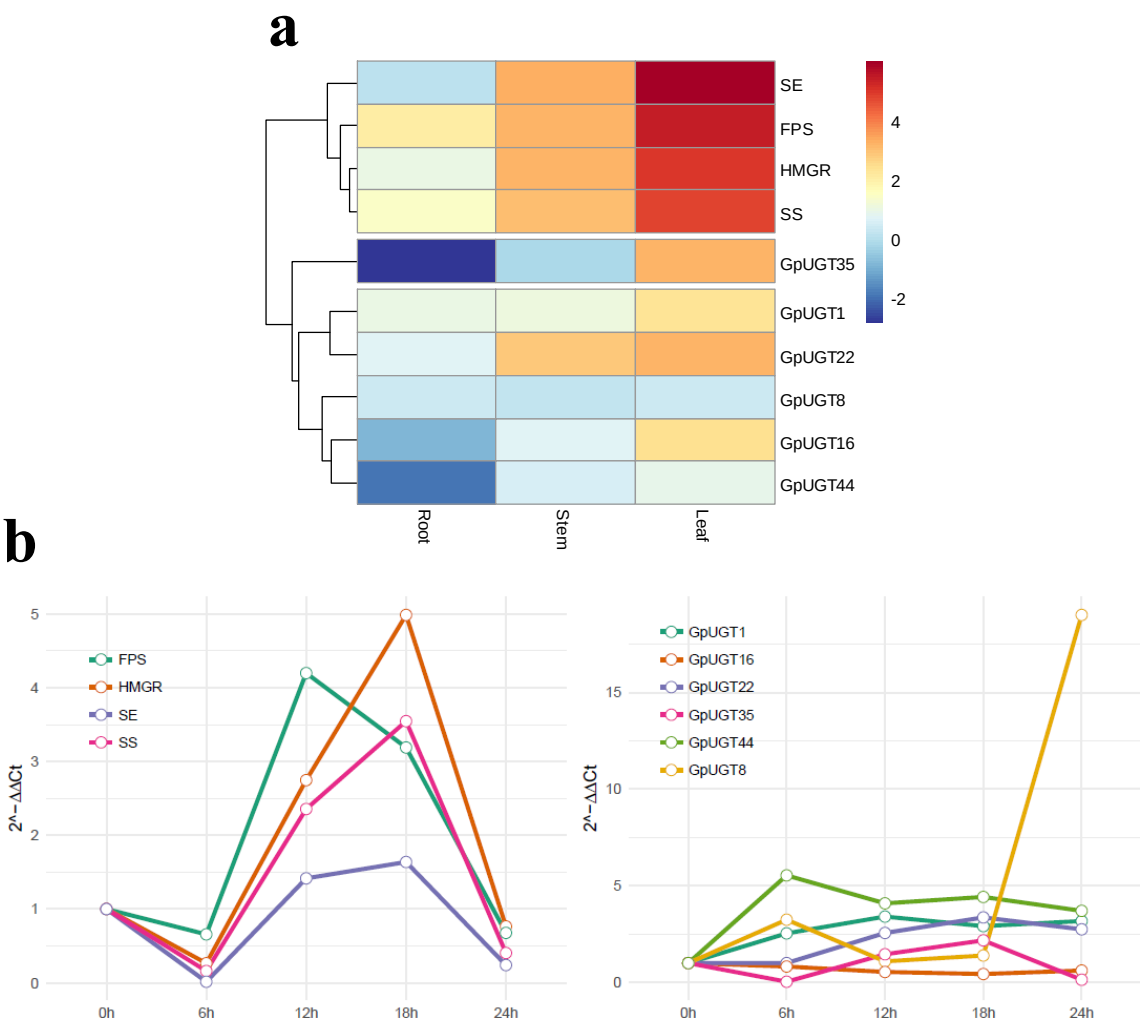

图 3.5 荧光定量 PCR 结果　　(a) 不同组织部位荧光定量 PCR 结果展示，所使用值为 $\log_e(2^{-\Delta\Delta Ct})$。(b) 不同诱导条件荧光定量 PCR 结果展示，横坐标为茉莉酸甲酯诱导时长，纵坐标为 $2^{-\Delta\Delta Ct}$。

4 讨论

4.1 绞股蓝皂苷合成途径预测

日本学者早期对于绞股蓝皂苷的分离揭示了其主要属于达玛烷皂苷的结构特点[18]，故而绞股蓝皂苷的合成途径可能与同属于达玛烷结构的人参二醇合成途径类似。通过亲缘关系分析，我们注意到 GpUGT35 属于 UGT94 家族，再结合人参二醇途径，我们预测 GpUGT35 的功能可能与绞股蓝皂苷VIII (人参皂苷 Rd)的合成有关(图 4.1)。

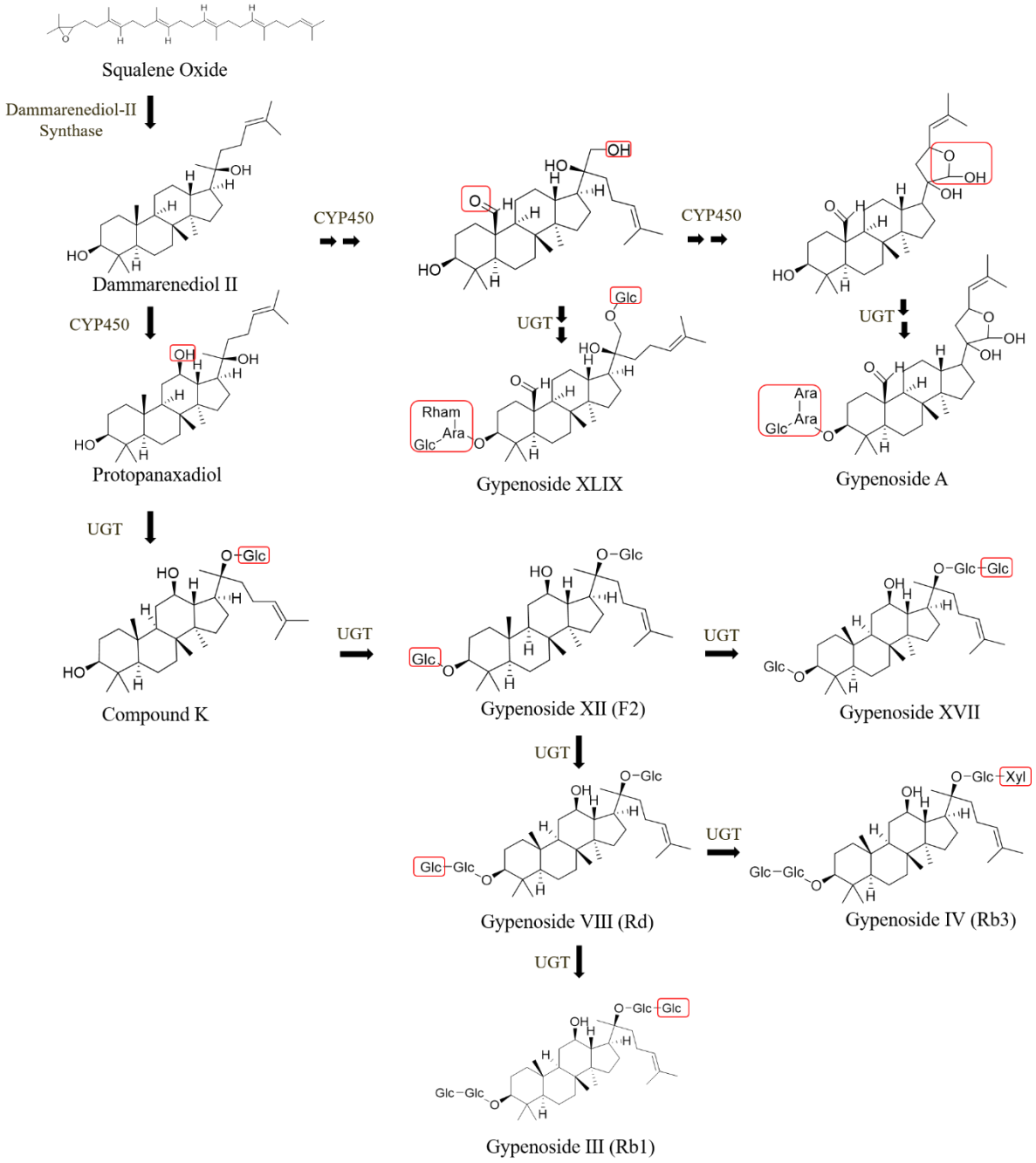

图 4.1 绞股蓝皂苷合成途径预测　　其中，两个连续的箭头代表多步反应；箭头上的标注为可能参与反应的酶；红色框中标注出了每一步反应后出现差异的基团。

在 MeJA 诱导处理中，我们发现在诱导后 6 小时上游基因以及 GpUGT35 的表达量都出现了下降，而这种趋势在二代数据和荧光定量 PCR 中均有存在。我们推测可能是由于诱导后，与 UGT 拥有相同底物的基因表达量升高，从而出现了竞争性抑制，也正是由于首先受到抑制，生物体内的负反馈机制导致了 6 小时之后的表达量升高补偿，在 18 小时处达到顶峰后，诱导效果渐渐消失，在 24 小时出逐渐降低(图 3.5b)。

本研究首次将二代与三代数据结合起来用于挖掘绞股蓝皂苷生物合成途径下游基因。并筛选到了一个可能与绞股蓝皂苷合成相关的糖基转移酶序列(GpUGT35)。对于绞股蓝皂苷的生物合成途径研究起到了一定的促进作用，为绞股蓝经济价值进一步的开发提供了理论支持。

4.2 三代测序技术优势

三代测序技术在转录组方面的优势，主要有以下几点：第一，长读长，PacBio 公司在酶以及荧光基团的连接位点上做出改进，使三代测序的读长最高可以大于 10kb，远远高于二代测序的 150bp；第二，测得转录本的准确性，三代测序测得转录本无需拼接也没有 GC 偏好性，而二代测序结果需要拼接，拼接时易出错，导致定量与结构分析不准确；第三，结构分析，由于直接测得转录本原始情况，可以更准确地定位剪切位点，从而进行 AS 与 APA 的分析。

在无参转录组方面，由于没有参考基因组，便准确定位剪切位点，所以结构分析方面不如有参转录组研究更准确，当然，也存在无参转录组的可变剪切分析，如国内学者对于矮牵牛以及无油樟的研究。

在本研究中，我们不仅发挥了三代测序长读长、无需拼接的优势，也结合了二代测序技术高通量、对于测序数据准确性高的优点。把两部分数据结合起来之后，我们可以得到更多的转录本以供分析之用，尽可能全地预测到所有可能的转录本。对于二代数据的直接引入，还在一定程度上弥补了三代测序无法测到过短序列 (小于 200bp) 的缺陷。尽管过短的序列最终没有纳入我们的选择范围，但是也为后续的转录组研究提供了尽可能全面的数据，尽量避免遗漏。

4.3 分析流程与一般流程的比较

一般公司给出的分析流程中，一般会对三代下机数据进行处理，之后使用二代数据进行矫正，最后进行下游分析。下游分析中包括功能注释、结构分析和差异富集分析。对于无参转录组而言，下游分析中除了一般分析中常见的功能注释以及差异富集分析之外，结构分析中仅包含了转录因子(Transcript Factor, TF)分析、长非编码 RNA

(long non-coding RNA, lncRNA)分析、简单重复序列 (simple sequence repeat, SSR) 分析。

本研究所使用的流程见图 3.1，可以看出，我们和公司使用的数据来源都是一样的，即二代和三代测序的原始数据文件，但是后续分析流程是不同的。这些公司给出的结果中，并不是所有的结果都可以出现在我们最后写成的文章中，或者说，大部分都不能出现。在一些特定方向的研究中，例如次生代谢途径解析，便仅用到了一部分注释结果，候选基因的挖掘分析都需要自己再去完善。

本研究的优势已于前文提起，但同时也存在不足之处。由于我们三代数据的数量并不是很好，将非全长校正过的 HQ 与 LQ 合并去冗余后仅有 8,550 条非冗余转录本，所以我们无法使用一般的分析流程继续往下分析。加入了二代序列拼接的转录本之后，所有的非冗余转录本数量达到了 140,157 条之多，大大扩充了参与下游分析的数据量。不幸的是，在扩充了数量的同时我们不但引入了冗余，也引入了拼接错误，也便造成在代谢途径分析中筛选候选目的基因的时候假阳性的上升。在次生代谢途径解析的过程中，尤其是功能验证的实验中，每一个假阳性都会造成很多不必要的浪费，所以如何挑全目的基因、如何去冗余、如何减少错误率是我们之后需要进一步研究的方向。一些可能的改进措施有，基因组测序、测序技术的进步以及开发新的算法等。

4.4 植物次生代谢中的糖基转移酶

GT Family 1，简称 GT1，是糖基转移酶的家族之一，属于 GT-B Fold，Inverting Clan。UDP 糖基转移酶是一类使用尿苷二磷酸修饰过的糖基作为底物的糖基转移酶。二者之间的关系界限并不清晰，最常见的表述就是：植物次生代谢相关的糖基转移酶大多都属于 UGT，而大多数植物 UGT 都属于 GT1。

问题起源于对于次生代谢途径相关 UGT 的挖掘中，我们注意到文献中对于参与植物次生代谢的糖基转移酶有描述为 UGT 的也有描述为 GT1 的，但是我们究竟应该挑选哪些作为我们的候选基因？或者说用哪个关键词筛选注释结果，才可以做到：挑选到所有参与植物次生代谢的糖基转移酶，既没有冗余，也没有遗漏？

经过对问题的进一步细化，我们发现，虽然难以确定二者之间的准确关系，但是我们可以通过验证已经进行功能验证的糖基转移酶所属分类来初步建立筛选方法。主要可以通过两步来进行判断：首先，我们需要明确文献中已报导的参与次生代谢的糖基转移酶有哪些；之后，进一步判断这些酶是否全是 GT1 或者全部都是 UGT 中的酶。第一步的相关工作主要通过文献调研来完成，寻找所有 1)文献报道过的 2)进行过功能验证的 3)参与植物次生代谢的糖基转移酶。第二步的相关工作主要通过数据库的注释

以及结果的统计来完成，在此工作中使用的两个数据库是包含了 CAZy 数据库的 dbCAN (http://cys.bios.niu.edu/dbCAN2/)以 Pfam (http://pfam.xfam.org/)。

在文献检索中，以一篇 2016 年的综述[6]为界，此文章已报导的之前发表的次生代谢相关的糖基转移酶从综述中获取，去除重复后共 113 条(UGT85H2, accession number: DQ875463, 在两篇文献中分别报导，记载了两次)，进行注释的共 101 条。其中，染色体序列 9 条，由于不好确认具体序列，暂不做研究；进行 Batch Entrez 时无法识别的有 7 条，3 条手动查询后补全入数据中，3 条，未能查询到蛋白序列(DY801582, FG404013, AK450655)，一条 ID 错误(AY14269)，从原文献校正后(AY142692)补全。之后发表的自行查阅文献获取，检索年份为 2015 年至今。检索结果中，共有 54 条记录(见附录表格 1)，其中，一条未提交，三条数据未公开。

由于所查询到的文献报道过的进行过功能验证的参与植物次生代谢的糖基转移酶，经过排除异常结果后，均同时属于 GT1 和 UDP-GT。因此如何从注释结果中挑选次生代谢相关的糖基转移酶便成为需要重点关注的问题。本研究使用以下策略：1) 使用所有的 Unigenes 进行 Pfam 注释和糖基转移酶注释；2) 将两部分注释结果取并集，在根据可能的长度去除过短的序列。这样挑选出来的序列最大限度地确保了所选序列的完整性，其中假阳性结果可以在后续筛选验证中加以去除。

5 结论

5.1 糖基转移酶的筛选方法：先使用所有的 Unigenes 进行 Pfam 注释和糖基转移酶注释；之后将两部分注释结果取并集，在根据可能的长度去除过短的序列；最后根据其他的分析结果例如进化树亲缘关系分析以及荧光定量 PCR 来对得到的初步结果进行进一步筛选，得到最后的结果。

5.2 基于上述流程，通过对注释结果的筛选，得到了 68 条 GpUGT 候选基因；结合聚类热图与进化树亲缘关系分析，筛选出了 6 条候选 GpUGT (GpUGT1, GpUGT8, GpUGT16, GpUGT22, GpUGT35, GpUGT44)；在荧光定量 PCR 验证中，GpUGT35 与上游基因表达趋势最为相近。据此，我们推测 GpUGT35 是最有可能参与绞股蓝皂苷生物合成的糖基转移酶序列。

参考文献

[1] Zhao Y, Zhou T, Li Z, et al. Characterization of Global Transcriptome Using Illumina Paired-End Sequencing and Development of EST-SSR Markers in Two Species of Gynostemma (Cucurbitaceae)[J]. Molecules, 2015,20(12):21214-21231.

[2] 范冬冬, 匡艳辉, 向世勰, 等. 绞股蓝化学成分及其药理活性研究进展[J]. 中国药学杂志, 2017(05):342-352.

[3] 史琳, 宋东平, 潘明佳, 等. 绞股蓝皂苷成分的研究进展[J]. 药物评价研究, 2011,34(6):456-464.

[4] 史琳, 王志成, 时圣明, 等. 绞股蓝皂苷水解产物化学成分和药理作用研究进展[J]. 药物评价研究, 2017(05):711-716.

[5] 郭溆, 罗红梅, 宋经元, 等. 糖基转移酶在植物次生代谢途径中的研究进展[J]. 2012.

[6] Costa V, Angelini C, De Feis I, et al. Uncovering the complexity of transcriptomes with RNA-Seq[J]. BioMed Research International, 2010,2010.

[7] Goodwin S, McPherson J D, McCombie W R. Coming of age: ten years of next-generation sequencing technologies[J]. Nature Reviews Genetics, 2016,17(6):333-351.

[8] Rhoads A, Au K F. PacBio Sequencing and Its Applications[J]. Genomics, Proteomics & Bioinformatics, 2015,13(5):278-289.

[9] 邹丽秋. 基于转录组分析的绞股蓝氧化鲨烯环化酶的挖掘与鉴定[D]. 北京协和医学院, 2017.

[10] Grabherr M G, Haas B J, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome[J]. Nature biotechnology, 2011,29(7):644.

[11] Haas B J, Papanicolaou A, Yassour M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis[J]. Nature protocols, 2013,8(8):1494.

[12] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data[J]. Bioinformatics, 2012,28(23):3150-3152.

[13] Tamura K, Stecher G, Peterson D, et al. MEGA6: molecular evolutionary genetics analysis version 6.0[J]. Molecular biology and evolution, 2013,30(12):2725-2729.

[14] Li B, Dewey C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC bioinformatics, 2011,12(1):323.

[15] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010,26(1):139-140.

[16] Kolde R. pheatmap: Pretty Heatmaps. R package version 1.0. 8[Z]. 2015.

[17] Razmovski-Naumovski V, Huang H W, Tran V H, et al. Chemistry and Pharmacology of Gynostemma pentaphyllum[J]. 2005,4(2-3):197-219.

[18] Tiwari P, Sangwan R S, Sangwan N S. Plant secondary metabolism linked glycosyltransferases: An update on expanding knowledge and scopes[J]. 2016,34(5):714-739.

[19] Zhao L, Ma X, Su P, et al. Cloning and characterization of a specific UDP-glycosyltransferase gene induced by DON and Fusarium graminearum[J]. 2018.

[20] Li J, Liang Q, Li C, et al. Comparative Transcriptome Analysis Identifies Putative Genes Involved in Dioscin Biosynthesis in Dioscorea zingiberensis[J]. 2018,23(2).

[21] He X, Zhao X, Gao L, et al. Isolation and Characterization of Key Genes that Promote Flavonoid Accumulation in Purple-leaf Tea (Camellia sinensis L.)[J]. 2018,8(1):130.

[22] Huang F C, Giri A, Daniilidis M, et al. Structural and Functional Analysis of UGT92G6 suggests Evolutionary Link between Mono- and Disaccharide Glycoside forming Transferases[J]. 2018.

[23] Mageroy M H, Jancsik S, Man Saint Yuen M, et al. A Conifer UDP-Sugar Dependent Glycosyltransferase Contributes to Acetophenone Metabolism and Defense against Insects[J]. 2017,175(2):641-651.

[24] Wilson A E, Feng X, Ono N N, et al. Characterization of a UGT84 Family Glycosyltransferase Provides New Insights into Substrate Binding and Reactivity of Galloylglucose Ester-Forming UGTs[J]. 2017,56(48):6389-6400.

[25] Kim O T, Jin M L, Lee D Y, et al. Characterization of the Asiatic Acid Glucosyltransferase, UGT73AH1, Involved in Asiaticoside Biosynthesis in Centella asiatica (L.) Urban[J]. 2017,18(12).

[26] Inoue S, Moriya T, Morita R, et al. Characterization of UDP-glucosyltransferase from Indigofera tinctoria[J]. 2017,121:226-233.

[27] Su X, Shen G, Di S, et al. Characterization of UGT716A1 as a Multi-substrate UDP:Flavonoid Glucosyltransferase Gene in Ginkgo biloba[J]. 2017,8:2085.

[28] Yin Q, Shen G, Di S, et al. Genome-Wide Identification and Functional Characterization of UDP-Glucosyltransferase Genes Involved in Flavonoid Biosynthesis in Glycine max[J]. 2017,58(9):1558-1572.

[29] Zhou K, Hu L, Li P, et al. Genome-wide identification of glycosyltransferases converting phloretin to

phloridzin in Malus species[J]. 2017,265:131-145.
[30] Chen H Y, Li X. Identification of a residue responsible for UDP-sugar donor selectivity of a dihydroxybenzoic acid glycosyltransferase from Arabidopsis natural accessions[J]. 2017,89(2):195-203.
[31] Zhou W, Bi H, Zhuang Y, et al. Production of Cinnamyl Alcohol Glucoside from Glucose in Escherichia coli[J]. 2017,65(10):2129-2135.
[32] Li P, Li Y J, Zhang F J, et al. The Arabidopsis UDP-glycosyltransferases UGT79B2 and UGT79B3, contribute to cold, salt and drought stress tolerance via modulating anthocyanin accumulation[J]. 2017,89(1):85-103.
[33] Zheng Y, Liao C, Zhao S, et al. The Glycosyltransferase QUA1 Regulates Chloroplast-Associated Calcium Signaling During Salt and Drought Stress in Arabidopsis[J]. 2017,58(2):329-341.
[34] Knoch E, Sugawara S, Mori T, et al. UGT79B31 is responsible for the final modification step of pollen-specific flavonoid biosynthesis in Petunia hybrida[J]. 2017.
[35] Xu G, Cai W, Gao W, et al. A novel glucuronosyltransferase has an unprecedented ability to catalyse continuous two-step glucuronosylation of glycyrrhetinic acid to yield glycyrrhizin[J]. 2016,212(1):123-135.
[36] Peng H, Yang T, Whitaker B D, et al. Calcium/calmodulin alleviates substrate inhibition in a strawberry UDP-glucosyltransferase involved in fruit anthocyanin biosynthesis[J]. 2016,16(1):197.
[37] Rojas Rodas F, Di S, Murai Y, et al. Cloning and characterization of soybean gene Fg1 encoding flavonol 3-O-glucoside/galactoside (1-->6) glucosyltransferase[J]. 2016,92(4-5):445-456.
[38] Smehilova M, Dobruskova J, Novak O, et al. Cytokinin-Specific Glycosyltransferases Possess Different Roles in Cytokinin Homeostasis Maintenance[J]. 2016,7:1264.
[39] Chen D, Sun L, Chen R, et al. Enzymatic Synthesis of Acylphloroglucinol 3-C-Glucosides from 2-O-Glucosides using a C-Glycosyltransferase from Mangifera indica[J]. 2016,22(17):5873-5877.
[40] Guo D D, Liu F, Tu Y H, et al. Expression Patterns of Three UGT Genes in Different Chemotype Safflower Lines and under MeJA Stimulus Revealed Their Potential Role in Flavonoid Biosynthesis[J]. 2016,11(7):e158159.
[41] Lu C, Zhao S, Wei G, et al. Functional regulation of ginsenoside biosynthesis by RNA interferences of a UDP-glycosyltransferase gene in Panax ginseng and Panax quinquefolius[J]. 2017,111:67-76.
[42] Yahyaa M, Davidovich-Rikanati R, Eyal Y, et al. Identification and characterization of UDP-glucose:Phloretin 4'-O-glycosyltransferase from Malus x domestica Borkh[J]. 2016,130:47-55.
[43] Cui L, Yao S, Dai X, et al. Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (Camellia sinensis)[J]. 2016,67(8):2285-2297.
[44] Blomstedt C K, O'Donnell N H, Bjarnholt N, et al. Metabolic consequences of knocking out UGT85B1, the gene encoding the glucosyltransferase required for synthesis of dhurrin in Sorghum bicolor (L. Moench)[J]. 2016,57(2):373-386.
[45] Dewitte G, Walmagh M, Diricks M, et al. Screening of recombinant glycosyltransferases reveals the broad acceptor specificity of stevia UGT-76G1[J]. 2016,233:49-55.
[46] De Bruyn F, De Paepe B, Maertens J, et al. Development of an in vivo glucosylation platform by coupling production to growth: Production of phenolic glucosides by a glycosyltransferase of Vitis vinifera[J]. 2015,112(8):1594-1603.
[47] Ahrazem O, Rubio-Moraga A, Trapero-Mozos A, et al. Ectopic expression of a stress-inducible glycosyltransferase from saffron enhances salt and oxidative stress tolerance in Arabidopsis while alters anchor root formation[J]. 2015,234:60-73.
[48] Song C, Gu L, Liu J, et al. Functional Characterization and Substrate Promiscuity of UGT71 Glycosyltransferases from Strawberry (Fragaria x ananassa)[J]. 2015,56(12):2478-2493.
[49] Ghose K, McCallum J, Sweeney-Nixon M, et al. Histidine 352 (His352) and tryptophan 355 (Trp355) are essential for flax UGT74S1 glucosylation activity toward secoisolariciresinol[J]. 2015,10(2):e116248.
[50] Funaki A, Waki T, Noguchi A, et al. Identification of a Highly Specific Isoflavone 7-O-glucosyltransferase in the soybean (Glycine max (L.) Merr.)[J]. 2015,56(8):1512-1520.
[51] Li W, Zhang F, Chang Y, et al. Nicotinate O-Glucosylation Is an Evolutionarily Metabolic Trait Important for Seed Germination under Stress Conditions in Arabidopsis thaliana[J]. 2015,27(7):1907-1924.
[52] Chen D, Chen R, Wang R, et al. Probing the Catalytic Promiscuity of a Regio- and Stereospecific C-Glycosyltransferase from Mangifera indica[J]. 2015,54(43):12678-12682.
[53] Liu Z, Yan J P, Li D K, et al. UDP-glucosyltransferase71c5, a major glucosyltransferase, mediates abscisic acid homeostasis in Arabidopsis[J]. 2015,167(4):1659-1670.

附录

表格 1 2015 年 1 月-2018 年 3 月 进行过功能鉴定的参与植物次生代谢的糖基转移酶汇总

ID	ID Type	Definition	Year	Reference
no submission	NA	TaUGT5	2018	[19]
MG488289	GenBank Nucleotide	Not Released	2018	[20]
MG488290	GenBank Nucleotide	Not Released	2018	[20]
ASA40331.1	GenBank	UDP-glucosyltransferase [Camellia sinensis]	2018	[21]
BAI22846.1	GenBank	UDP-sugar flavonoid glycosyltransferase [Vitis vinifera]	2018	[22]
ASU43997.1	GenBank	UDP-glycosyltransferase UGT5b [Picea glauca]	2017	[23]
ANN02875.1	GenBank	UGT84A23 [Punica granatum]	2017	[24]
AUR26623.1	GenBank	UDP-glucosyltransferase 73AH1 [Centella asiatica]	2017	[25]
BBB16127.1	GenBank	UDP-glucosyltransferase [Indigofera tinctoria]	2017	[26]
KX371617	GenBank Nucleotide	Not Released	2017	[27]
NP_001279020.1	NCBI Reference Sequence	isoflavone 7-O-glucosyltransferase 1-like [Glycine max]	2017	[28]
D3UAG5.1	UniProtKB/Swiss-Prot	AltName: Full=UDP-glycosyltransferase 88F1	2017	[29]
D3UAG4.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 88F4	2017	[29]
Q9LZD8.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 89A2	2017	[30]
AAS55083.1	GenBank	UDP-glucose glucosyltransferase [Rhodiola sachalinensis]	2017	[31]
OAP09184.1	GenBank	UGT73C5 [Arabidopsis thaliana]	2017	[31]
Q9T080.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 79B2	2017	[32]

ID	ID Type	Definition	Year	Reference
Q9T081.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 79B3	2017	[32]
KYP54814.1	GenBank	Glycosyltransferase QUASIMODO1 [Cajanus cajan]	2017	[33]
AAD55985.1	GenBank	UDP-galactose:flavonol 3-O-galactosyltransferase [Petunia x hybrida]	2017	[34]
ANJ03631.1	GenBank	UDP-glycosyltransferase [Glycyrrhiza uralensis]	2016	[35]
AJW28718.1	GenBank	UDP-glucosyltransferase [Fragaria vesca subsp. vesca]	2016	[36]
NP_001345940.1	NCBI Reference Sequence	flavonol-3-O-glucoside/galactoside (1->6) glucosyltransferase [Glycine max]	2016	[37]
OAO89564.1	GenBank	UGT76C1 [Arabidopsis thaliana]	2016	[38]
OAO93987.1	GenBank	UGT76C2 [Arabidopsis thaliana]	2016	[38]
OAP13723.1	GenBank	UGT85A1 [Arabidopsis thaliana]	2016	[38]
AMM73095.1	GenBank	C-glycosyltransferase [Mangifera indica]	2016	[39]
ANW09827.1	GenBank	UDP-glycoslytransferase 3 [Carthamus tinctorius]	2016	[40]
ANW09829.1	GenBank	UDP-glycosyltransferase 25 [Carthamus tinctorius]	2016	[40]
ANW09828.1	GenBank	UDP-glycosyltransferase 16 [Carthamus tinctorius]	2016	[40]
ALE15280.1	GenBank	UDP-glycosyltransferase 3GT2 [Panax quinquefolius]	2016	[41]
NP_001315912.1	NCBI Reference Sequence	crocetin glucosyltransferase, chloroplastic-like [Malus domestica]	2016	[42]
ALO19890.1	GenBank	UDP-glycosyltransferase 84A22 [Camellia sinensis]	2016	[43]
ALO19888.1	GenBank	UDP-glycosyltransferase 78A14 [Camellia sinensis]	2016	[43]
ALO19889.1	GenBank	UDP-glycosyltransferase 78A15 [Camellia sinensis]	2016	[43]
Q9SBL1.1	UniProtKB/Swiss-Prot	RecName: Full=Cyanohydrin beta-glucosyltransferase	2016	[44]

ID	ID Type	Definition	Year	Reference
AGL95113.1	GenBank	UDP-glycosyltransferase 76G1 [Stevia rebaudiana]	2016	[45]
AEW31188.1	GenBank	glucosyltransferase [Vitis vinifera]	2015	[46]
AIF76152.1	GenBank	UDP-glucosyltransferase UGT85U1 [Crocus sativus]	2015	[47]
AIF76151.1	GenBank	UDP-glucosyltransferase UGT85U2, partial [Crocus sativus]	2015	[47]
AIF76150.1	GenBank	UDP-glucosyltransferase UGT85V1 [Crocus sativus]	2015	[47]
XP_004307485.1	NCBI Reference Sequence	PREDICTED: crocetin glucosyltransferase, chloroplastic-like [Fragaria vesca subsp. vesca]	2015	[48]
XP_011468178.1	NCBI Reference Sequence	PREDICTED: anthocyanidin 3-O-glucosyltransferase 2-like [Fragariavesca subsp. vesca]	2015	[48]
XP_004303953.1	NCBI Reference Sequence	PREDICTED: putative UDP-glucose flavonoid 3-O-glucosyltransferase 3 [Fragaria vesca subsp. vesca]	2015	[48]
XP_004303954.2	NCBI Reference Sequence	PREDICTED: putative UDP-glucose flavonoid 3-O-glucosyltransferase 3 [Fragaria vesca subsp. vesca]	2015	[48]
XP_004303955.1	NCBI Reference Sequence	PREDICTED: UDP-glucose flavonoid 3-O-glucosyltransferase 6-like [Fragaria vesca subsp. vesca]	2015	[48]
AGD95005.1	GenBank	lignan glucosyltransferase [Linum usitatissimum]	2015	[49]
NP_001304440.2	NCBI Reference Sequence	isoflavone 7-O-glucosyltransferase UGT4 [Glycine max]	2015	[50]
OAP07463.1	GenBank	UGT74F2 [Arabidopsis thaliana]	2015	[51]
Q9FI98.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 76C4	2015	[51]
Q9FI97.1	UniProtKB/Swiss-Prot	RecName: Full=UDP-glycosyltransferase 76C5	2015	[51]
A0A0M4KE44.1	UniProtKB/Swiss-Prot	Full=UDP-glycosyltransferase 13; Short=MiUGT13; AltName:Full=C-	2015	[52]

ID	ID Type	Definition	Year	Reference
OAP14418.1	GenBank	glycosyltransferase; Short=MiCGT; UGT71C5 [Arabidopsis thaliana]	2015	[53]

综述

全长转录组研究进展

摘要：自 2016 年玉米和高粱两篇全长转录组的文章在 Nature Communication 上发表以来，全长转录组测序逐渐发展起来，不同于这两篇以及后续的文章所采用的 PacBio RS2 平台，目前全长转录组测序主要使用第三代 Sequel 平台完成。本文基于最新的 Sequel 平台，简要介绍研究思路、测序原理、标准化流程、下游数据挖掘，并对在无参转录组中存在的去冗余问题提出了一种参考解决方案。以期能为想要深入此领域的研究人员提供参考，拓宽研究思路。

关键词：全长转录组；PacBio Sequel；无参转录组去冗余

1 测序发展概要

从第一代 Sanger 测序开始，测序技术的发展共历经三代。他们分别是：第一代 Sanger 测序主打高精度，方法有二，分别是：Sanger 与 Coulson 发明的双脱氧链终止法和 Maxam 与 Gilbert 发明的化学裂解法。1987 年基于荧光标记的 Sanger 测序仪的发明，使得前者在时间的筛选中得以保留；第二代测序技术 NGS, Next Generation Sequencing 或称边合成边测序(SBS, Sequencing By Synthesis) 主打高通量，方法有三，分别是：Ronaghi 与 Nyren 的焦磷酸测序、荧光标记多聚核苷酸与模板互补配对和荧光标记的脱氧核苷酸分步反应。现在的测序巨头 Illumina 采用的是第三种方法，建库核心是桥式 PCR，以用来放大荧光信号。第三代测序技术单分子测序主打长读长，方法有三，两种是荧光信号测序，分别是 Helicos Biosciences 的 True Single-Molecule Sequencing (tSMS) Heliscope Sequencing（已被淘汰）、Pacific Biosciences 的单分子实时测序技术(SMRT, Single Molecule Real Time Sequencing) 以及一种电信号测序，Oxford Nanopore Technologies 的纳米孔(Nanopore)测序技术。PacBio 的核心是零模波导孔(ZMW, Zero Mode Wave Hole)，而 nanopore 的核心是名为 ratchet 的酶[1]。

对于全长转录组(Iso-Seq, Isoform Sequencing)来说，主要采用的是 PacBio 的 SMRT 技术，基于此技术推出的测序平台主要有三代，分别是 RS，RS2 和 Sequel。最小的测序单位是一个 ZMW 孔，而实际操作过程中，则是一个 SMRT Cell，在这个 SMRT Cell 中有着许多的 ZMW 孔，并且其数量正在随着不断进步的技术而增加。在测序时，每一个 ZMW 中只有一个酶，其荧光信号是由合成反应释放的，荧光基团连接在磷酸末端上，

注意这里没有连接在碱基上的荧光基团是第三代测序技术的长读长的主要原因之一。在 DNA 测序中，反应的荧光基团可以通过信号捕获器捕捉，而在反应中因为原始的序列仅仅经过末端修复而未大量 PCR 建库，所以上面的甲基化是可以被测出的，但是 RNA 中因为要经过反转录 PCR 合成 cDNA 以及 PCR 扩增建库，所以测到的甲基化信息已经不再准确。具体的建库过程会在下文中详细描述。

2 研究思路

对于全长转录组来说，主要的分类从应用对象可以分为动物、植物、以及人类（医学方向），从有无参考基因组角度可以分为有参和无参两种。

在没有参考基因组的时候，我们对于校正去冗余后的数据主要可以进行注释、定量与差异分析富集（建立在二代数据基础之上）、lncRNA 与 CDS 预测。主要的应用方向有：

1) 获得全长参考序列，三代测序技术的长读长，可以使我们不用组装，便可以得到原始的转录本(Isoform)信息，再依此进行的定量，相比于二代组装后的 Unigene 来说便有了更高的准确性，从而更加高效准确地寻找差异基因。
2) 结构性分析，现主要可以使用 PacBio 官方推荐的 Cogent 软件，对于所有转录本的 exon 区域片段组装后形成参考基因组，再使用 Iso-Seq 测序数据与其比对，进行结构分析，如可变剪切 (AS, Alternative Splicing) 。
3) 功能性分析，先对数据定量之后进行差异基因分析，再从注释结果之中进行差异基因富集，得到可能参与该功能的代谢通路和 GO 三大水平的功能[2]。
4) 差异转录组，主要用于同一物种表型差异的分析，挖掘与表型差异相关的转录本，进一步分析其有别之处。
5) 比较转录组，主要用于近缘物种亲缘关系、及不同物种及其亚种之间的差异转录本分析[3]。

一旦我们拥有了参考基因组，一些不确定性便可以规避，我们主要对校正后的数据进行与参考基因组比对、新基因与新转录本预测、新基因与未比对上转录本注释、可变剪切(AS)分析、可变多聚腺苷酸化(APA, Alternative Ploy-A)分析、融合基因分析、lncRNA 预测、转录因子(TF)预测以及后续差异分析及功能富集（同样建立在二代测序数据基础之上）。主要的应用方向有：

1) 完善转录组，主要为上述的 AS 至 TF 的分析。
2) 完善基因组，完善基因组注释结果[4]。

3) 功能性研究，如果有后续的差异基因和功能富集，便可以进一步分析相关功能与富集到相应通路转录本的联系。

4) 差异转录组，研究同一物种不同处理之间的或不同表型之间的差异转录本[5]。

5) 比较转录组，同无参中描述[6]。

6) 动态转录组，比较同一物种不同时间中，转录本随发育阶段的变化情况[7]。

3 标准化流程

3.1 样品准备

送测 RNA 样品的种类主要有，动物组织、植物组织、培养细胞、全血及菌体。在判定样品是否合格的过程中，主要使用了四步检测，分别是：Nanodrop 检测、凝胶电泳检测、Qubit 检测以及 Agilent 2100 生物分析仪检测。判定合格主要有五点标准，分别是：浓度及总量(total RNA) 大于等于 300 纳克每微升，总量要达到 5 微克以上；OD260/280 要在 2.0-2.2 之间；OD260/230 要在 1.2-2.1 之间；最重要的是 RIN 值，需要大于 8，如果是无 28S 的物种，无法测得 RIN 值，需要 2100 检测基线平稳；其他条件如无颜色、无不溶性杂质等，也需要考虑在内。如需更详细的参数，请询问测序公司销售，不再赘述。

3.2 文库制备

样品填写好实验分组表格后，寄往测序公司。在测序公司收到样品后，首先对收到的样品进行编号，以便在公司内部进行流通，之后由 RNA 提取人员提取 RNA，由质检部门负责提取 RNA 的质控，再由建库部门进行建库，最后上机测序。

在建库策略方面，现行 Sequel 平台主要的建库策略是：不筛选片段建库加 4k 以上大片段建库，之后二者等摩尔量混合以构建混合文库，用于上机测序。这里要注意的是，在此前的 PacBio 测序平台中，例如 RS 以及 RS2 之中，都是采用 1-2kb、2-3kb、>3kb 或是类似的方法分段建库测序的，之所以改成现在这样是因为在 Sequel 的升级中，ZMW 的孔深度有所增加，减小了对于长度较短的 SMRTbell 偏好性，而 4kb 以上的大片段在转录本中所占比例并不是非常多，为了避免遗漏长片段，故而对于 4k 大片段进行富集建库并等摩尔量混合共同上机测序。

在建库流程上，首先富集带有 Poly-A 的 mRNA 片段，使用带有 5’接头和 3’接头的引物进行反转录 PCR 使之成第一链 cDNA，之后 PCR 至建库浓度；筛选出 4k 以上大片段，和不筛选的片段等摩尔量混合（可选）；之后连接 SMRTbell adaptor 使之成为 SMRTbell，此步又分为接头连接、模板纯化、引物结合、聚合酶绑定，最后上机测序。

3.3 上机测序

测序的原理在前文有所述及，下面主要讲述上样流程以及测序过程。

上样流程，首先将 MagBead[8]磁珠与建好的文库混合，这是一种表面固定着 Oligo dT 的磁珠。之后，SMRTbell 上的 Poly-A tail 会与磁柱上的 Oligo dT 结合，结合上的 SMRTbell 可以分为三种，第一种是插入片段过短的、第二种是长度适中的，真正要被测序的、第三种是引物聚合体。将 MagBead 与 SMRTbell 复合体加入 SMRT cell 中，当磁珠在 cell 底部转动时，也会带着 SMRTbell 经过 ZMW 孔，其中，第一种和第三种聚合体都不会触及 ZMW 底部，从而无法停留在 ZMW 中，而第二种长度足够（这个最小长度是 200bp，如果长度再短可以直接考虑使用二代的数据弥补），酶被固定在底部，AT 之间的氢键被扯断，从而使整个 SMRTbell 停留在 ZMW 当中，进行后续测序。值得注意的是，ZMW 的大小十分有讲究，不但长度可以使 SMRTbell 一端的酶结合在底部，其直径约 70nm，小于荧光信号波长，从而使测序信号限定在 ZMW 的孔中，尽量减少对于其他孔的干扰，从而保证测序精度。

测序时，固定在底部的 DNA 聚合酶将反应体系中的荧光标记 dNTP 聚合到模板上，这些荧光标记的位点是在磷酸末端，每一次合成都会释放出相应的荧光，这些荧光局限在底部 20-30nm 的高度内，在孔的上开口处信号便十分微弱，不会对其他孔产生很大影响。光信号捕获器捕获合成时释放的荧光，通过碱基信号识别 base calling，转化为 ATCG 的字符串信息，存储下来，成为下机数据。对后续分析有用的下机数据是

3.4 信息分析

3.4.1 下机数据处理

下机数据首先要经过 SMRT Link 软件套装来处理，现行版本主要是 5.1，这是一款官方给出的软件，可以自由下载使用，所有的 PacBio 下机数据处理首先都要经过它的处理(https://github.com/PacificBiosciences/SMRT-Link)。

测序时，每一个 ZMW 孔中产生的所有测序信号叫做 subreads，注意这个 subreads 可能有全长(FL, Full Length)，也可能有测到一半中止的，也就是非全长(nFL, non Full Length)的，事实上测到最后一条最后一个碱基正好读取了整数圈的情况不会太多。这些 subreads 通过去除接头，分为若干个片段，进行孔内自校正，最后得到的一致性序列称为环化一致性序列(CCS, Circular Consensus Reads) 这个环化一致性序列可以认为是与之前的 PacBio 平台中的 Read of Insert 是同样的东西，二者之间仅在名称上有所改变(https://github.com/PacificBiosciences/SMRT-Link/wiki)。

在完成了自身校正后，通过是否为全长的判断标准将所得到的 CCS 序列分为全长和非全长两类，这个判断标准有以下三点：5’UTR、Poly-A、3’UTR。只有同时拥有这三个部分的 CCS 才是全长的。其中，全场的 CCS 使用 ICE （属于 SMRT Link）软件进行孔间聚类，之后，将非全长的 CCS 与经过 ICE 的 CCS 一同送入 Arrow 软件中（在之前的版本中也叫做 Quiver，同样仅在名称上有所改变，所完成的工作基本一致）nFL 对 FL 的 CCS 进行校正，nFL 进行校正后并不参与后续分析，此处尤其值得注意，全长非嵌合序列(FLNC, Full Length and Non Chimeric)中的 HQ 和 LQ 只是依据提供支持的 CCS 数量多少划分的，而所有的 FLNC 都是 FL 的。

3.4.2 二代数据校正

一般来说，现在的三代全长转录组数据都会推荐同时进行二代测序，一是因为二代数据的直接测序精度高，三代数据的直接测序精度有在 85%左右，再进行了孔内自校正之后，这个准确度可以提升约 10%，再通过 nFL 校正，可以提高 2-3%，最后使用二代数据校正，可以再提高 1-2%；二是因为二代数据的高通量，三代数据通量也不低，但是成本十分高昂，要达到和二代一样的覆盖度(Coverage)的时候，价格相比于二代测序高出很多，从下机数据的饱和曲线来看，仅用三代的数据曲线还未达到饱和（趋势趋于平缓），这时使用二代准确的高通量 reads 可以弥补数据量的不足，达到较好的覆盖度；三是二代数据可以用来定量，三代数据测序深度低，数据量少，用来定量不准确，使用二代数据定量便可以进行后续的差异分析。

在进行二代校正的时候，当前主要推荐的软件是 LoRDEC[9]，本软件主要用于使用二代数据对三代数据进行校正。经过比较，优于 proovread[10]，而 LSC[11]软件主要用于人类数据的校正。

3.4.3 有参全长转录组后续分析

有参因为有参考基因组，便可以与参考基因组比对，同时使用参考基因组对于测序数据予以修正，进一步提高准确性，在比对之后，便是新基因与新转录本预测、新基因与未比对上转录本注释、以及后续差异分析（回帖目标使用补全新基因后的基因组）及功能富集这些流程化的工作。详细阐述会写在数据挖掘中。

3.4.4 无参全长转录组后续分析

无参因为没有参考基因组，便无从比对，所以进行聚类去冗余，生成 Unigene 之后，全部进行注释。再用二代数据对 Unigene 进行回帖定量、差异分析、功能富集。同样，这并不能最终拿到我们想要的基因，所以也需要进一步挖掘数据。

4 数据挖掘

因为进一步的数据挖掘大多是根据具体情况具体分析，每个课题所想要筛选的目的基因也不同，后续的挖掘便十分个性化，有些分析用得到，有些用不到，虽然下面一些分析也被一些公司做到了流程之中，但是当我们真正想要使用相关数据的时候，很多还要我们重新提取并整理数据，然后重新绘制图片。

4.1 结构分析

在二代转录组测序中，由于测序数据需要拼接才可以得到转录本，而拼接出的转录本却不一定准确。但是三代测序不同，其长读长的优势可以直接获取整条转录本，从而准确的预测在转录时发生的结构变化时间，例如 AS 以及 APA，融合基因等，lncRNA 预测主要是通过三种软件分析以及 Pfam 注释完成，而转录因子的分析基本可以归到注释里面，因为从分析方法上来说还是通过数据库注释完成的。

在有参结构分析中，不得不提的就是 Tapis 流程[5]，此流程最先报导在 2016 年发表的对于高粱的研究工作中，目前最新版是 TAPIS 1.2.1，可以自动与参考基因组比对，根据基因组校正转录本，发现新基因和新转录本，分析 APA 和 AS。其中的 AS 分析可以使用 SUPPA[12]来替换，分析类型更多更全。融合基因使用 cDNA_Cupcake 中 Tofu 的 find_fusion.py[13]可以完成。lncRNA 分析可以通过 CPC[14]、CNCI[15]、PLEK[16]以及与 Pfam 比对来完成。

对于无参结构分析中，诺禾致源公司提出了 AS 分析，主要是使用 Cogent[17]软件配合 SUPPA 来完成，但是准确率有待提升。

4.2 差异分析

在差异分析中，我们首先要认识到在做差异分析之前要先定量，定量使用的是二代数据。对于有参中，有生物学重复的可以使用 DESeq2[18]来完成，无生物学重复的可以使用 DGESeq[19]，对于无参，可以使用 RSEM[20]来完成。

4.3 深入挖掘

在进行差异分析后，我们会拿到很多差异表达基因，甚至，在无参中里面还存在一些冗余，即一些相似度很高的转录本干扰着我们的分析，去冗余这一部分会在讨论中进一步描述。

对于有参的测序数据和不考虑冗余的无参数据来说，在差异分析后，我们主要的缩小研究范围的方法是差异基因富集，主要涉及的两个数据库是 GO[21]和 KEGG[22]，一般在公司给出的结题报告中我们会拿到所有差异表达基因的功能富集结果，此时，我

们就应该通过在实验设计之初就想好的目的来寻找和实验差异处理目的相关的功能和通路，从中把富集到的基因或转录本提取出来，进行下一步分析。也可以通过进一步注释、共表达分析、蛋白互作网络、进化树分析等方法，来进一步缩小范围

在确定了一部分转录本和目标功能相关后，我们主要进行定性验证和功能验证两个步骤来最终锁定目的基因。当然，如果锁定的过多，验证起来也十分麻烦，在这里仅给出参考数字为数十个到一百以内，为项目经验数值，仅供参考。

定性验证主要方法有：荧光定量核酸扩增检测 (qPCR, Real-time Quantitative PCR Detecting System) 、Northern Blot、 荧光原位杂交组织化学法 FISH 等，其中，当属 qPCR 最常用。

功能验证方法主要有两大类，一类是过表达，我们可以过表达某个基因，方法如构建过表达转基因载体，侵染细胞或愈伤组织，观察过表达后是否现象更为明显。一类是减少表达敲除或者敲减某个基因的表达，方法如 siRNA 干扰，CRISPR/Cas9 切除等观察敲除敲减后是否现象反转或不明显。后续如果有能力，可以寻找相应蛋白的底物和产物，明确其作用并且预测其等电点等属性。

5 技术缺陷与应对

三代测序主要强调几点优势，除去在基因组测序中强调的直接测得甲基化信息之外，在转录组方面值得提起的优势便是可以测得全长转录本无需组装以提高后续分析精度，后续分析又可分为结构分析和定量之后的差异分析。但是三代测序也有其不可避免的缺陷。让我们从一个理想的测序开始谈起。

所谓理想测序，就是在测序中，所获取的数据有着高还原性（如还原甲基化信息），高准确度和无限读长。这个时候，测序数据的测序深度也许便不再重要，因为 1×的数据便可能覆盖了整个基因组。但显然，现在的测序还远远达不到这一点，高准确度的通量不够，高通量的读长不够，长读长的通量和准确度都有待提高。

便如前文所提到的三加二策略，现在人们的解决方案主要是将不同优势的测序技术应用到不同的领域，在优势可以互补的地方相互结合。但是结合后又有新的问题出现，有参的测序有着基因组，我们可以使用基因组数据来校正三代数据。但是无参转录组测序中，我们没有参考基因组去比对，这就造成了许多冗余的转录本无法被去除。在后续的差异分析和验证时我们便会发现有很多相似的转录本，在验证时数量过多而使得实验人员无从下手验证。

说起去冗余，就要从第一步的 SMRT Link 说起，在 SMRT Link 中 ICE 首先会将相

似的 CCS 聚类，之后，在二代数据校正之后我们会使用 CD-HIT[23]去冗余，但是需要注意的是，这里的相似性阈值都十分高，-c 参数代表了序列相似性阈值，高于这一参数的两条转录本会留下其中较长的一条。这一参数在公司流程中被设定为 -c = 0.99，便导致有很多很相似的冗余是去不掉的，下面描述一种应对方法，仅供参考。

问题的出现：首先，我们三代的测序数据并不是很多，所以我们使用二代测序数据拼接后的结果来补充我们的结果（当时在补充的时候没有想到的是，这样的补充正好可以填补三代测序技术中 ZMW 无法测到 200bp 一下片段的缺陷），在将两份数据合并之后，我们便有了比一般三代数据还要多的冗余，所以去冗余便成为了一项亟待解决的问题。

应对方案：首先我们的目的基因是萜类合成中关键的两个酶，一个是 CYP450，一个是 UGT，此处以 P450 为例。我们将 CD-HIT 以 -c = 0.85 的参数运行，出来的结果我们命名为 cd.out ，在将 cd.out 翻译成为蛋白序列之后我们重新注释结果，我们使用了 Pfam 和 Swiss-Prot 两个数据库的结果来提取注释到 P450 的结果，后面我们经过了三步判断流程以筛选去冗余，它们分别是：

1) 95%相似性判断，低于 95%相似性的序列直接送入下游分析，高于 95%的序列进行预筛选；
2) 99%相似性判断，低于 99%相似性的序列聚类后送入人工筛选，高于 99%的序列保留其一，送入下游分析；
3) 人工筛选，标准如下：
    a) 氨基酸长度大于 300；
    b) 相同 cluster 中序列比对：
        i) 尽量保留长序列，每个 cluster 保留 1-2 条；
        ii) 如果有 AS，两条全部保留；
        iii) 如果数条短序列被一条长序列完全覆盖，留下长序列；
        iv) 对于单碱基错误，应属于 SNP，在有多条一致的情况下，舍去不同的哪一条
    c) 对于难以确认如何保留的，进行 BLAST 以及 CDS 预测，保留比对上同源物种，相似功能基因相似性高的，CDS 区域长的。
4) 下游分析，
    a) 使用文献报道的上游基因，与分析到的序列做进化树，聚到同一支上的说明有着较高的同源性，更有可能编码目的蛋白。

b) 也可以通过表达量 Heatmap 聚类，与上游基因表达模式相似的，更有可能编码目的蛋白。

6 发展趋势

现在的测序界是三代技术共存的，第一代测序主要用于分子克隆中的载体上的目的基因测序，由于不需要高通量，只注重精确度，Sanger 测序很好地胜任此工作。第二代测序由于高通量以及 150bp 及以下的读长及其日益降低的成本，使得在一些不需要长读长，注重高通量的测序任务中，将二代测序视为首选，并且一些本来就很短的转录本，例如细菌、微生物的测序，在三代测序之中是无法测到的。三代测序由于长读长，在减少组装带来的误差、提高长片段精确度方面有着很大的贡献，但是无疑成本高昂是其缺陷之一。

PacBio 的 SMRT Sequel 平台还在不断改进之中，成本降低是必然的，但是降多少和降低的速度就只能拭目以待了，每一个 SMRT cell 中的产出数据量还会提升，试剂和酶也会升级，这就会使得 PacBio 在通量增加的基础上，测序长度也有可能进一步提升。三代数据对于转录组，尤其是无参转录组来说，准确度是二代无法比拟的，二代的数据虽然直接测序的准确性高，但是测出来的数据却需要拼接，便如将一本书粉碎后在拼起来，当然是粉碎片段大的拼起来难度较低，而三代数据理论上可以直接获取全长转录本而不用拼接，便大大提高了后续分析的准确性。所以可以说，在转录组测序方面，三代测序是优于二代测序的。三代测序技术在全长转录组测序上的应用也会越来越多。

在未来的发展中，PacBio 平台还将继续改进，三代测序所需成本将会有所下降。Nanopore 由于其便携性，以及其不同以往的测序机制，或许会让测序仪更加普及，而不仅仅将测序工作局限在测序公司和实验室当中。将 Nanopore 的测序数据结合物联网(IoT, Internet of Things)，将会极大地提升测序数据库中的数据量[1]。在将机器学习算法应用于疾病分析中，相信对于疑难疾病的治疗将会有新的解决方案。

综述参考文献

[1] Shendure J, Balasubramanian S, Church G M, et al. DNA sequencing at 40: past, present and future[J]. Nature, 2017,550(7676):345-353.

[2] Xu Q, Zhu J, Zhao S, et al. Transcriptome Profiling Using Single-Molecule Direct RNA Sequencing Approach for In-depth Understanding of Genes in Secondary Metabolism Pathways of Camellia sinensis[J]. Frontiers in Plant Science, 2017,8:1205.

[3] Workman R E, Myrka A M, Tseng E, et al. Single molecule, full-length transcript sequencing provides insight into the extreme metabolism of ruby-throated hummingbird Archilochus colubris[J]. GigaScience, 2018:y9.

[4] Wang B, Tseng E, Regulski M, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing[J]. Nature Communications, 2016,7:11708.

[5] Abdel-Ghany S E, Hamilton M, Jacobi J L, et al. A survey of the sorghum transcriptome using single-molecule long reads[J]. Nature Communications, 2016,7:11706.

[6] Zhu F, Chen M, Ye N, et al. Proteogenomic analysis reveals alternative splicing and translation as part of the abscisic acid response in Arabidopsis seedlings[J]. The Plant Journal, 2017,91(3):518-533.

[7] Kuang Z, Boeke J D, Canzar S. The dynamic landscape of fission yeast meiosis alternative-splice isoforms[J]. Genome research, 2017,27(1):145-156.

[8] Raley C, Munroe D, Jones K, et al. Preparation of next-generation DNA sequencing libraries from ultra-low amounts of input DNA: Application to single-molecule, real-time (SMRT) sequencing on the Pacific Biosciences RS II.[J]. bioRxiv, 2014.

[9] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction[J]. Bioinformatics, 2014,30(24):3506-3514.

[10] Hackl T, Hedrich R, Schultz J, et al. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus[J]. Bioinformatics, 2014,30(21):3004-3011.

[11] Au K F, Underwood J G, Lee L, et al. Improving PacBio long read accuracy by short read alignment[J]. PloS one, 2012,7(10):e46679.

[12] Trincado J L, Entizne J C, Hysenaj G, et al. SUPPA2 provides fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions[J]. bioRxiv, 2017:86876.

[13] Wang M, Wang P, Liang F, et al. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation[J]. New Phytologist, 2018,217(1):163-178.

[14] Kong L, Zhang Y, Ye Z, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine[J]. Nucleic acids research, 2007,35(suppl_2):W345-W349.

[15] Sun L, Luo H, Bu D, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts[J]. Nucleic acids research, 2013,41(17):e166.

[16] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme[J]. BMC bioinformatics, 2014,15(1):311.

[17] Li J, Harata-Lee Y, Denton M D, et al. Long read reference genome-free reconstruction of a full-length transcriptome from Astragalus membranaceus reveals transcript variants involved in bioactive compound biosynthesis[J]. Cell discovery, 2017,3:17031.

[18] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for

RNA-seq data with DESeq2[J]. Genome biology, 2014,15(12):550.
[19] Wang L, Feng Z, Wang X, et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data[J]. Bioinformatics, 2009,26(1):136-138.
[20] Li B, Dewey C N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome[J]. BMC bioinformatics, 2011,12(1):323.
[21] Gene O C. The Gene Ontology (GO) database and informatics resource[J]. Nucleic acids research, 2004,32(suppl_1):D258-D261.
[22] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000,28(1):27-30.
[23] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data[J]. Bioinformatics, 2012,28(23):3150-3152.

译文

DNA 测序 40 年：过去，现在与未来

摘要: 谨以此篇综述纪念 DNA 测序 40 周年，在这 40 年间，我们见证了众多科技的革命的出现，见证了测序技术从几千个碱基到第一个人类基因组，再到百万人的基因组测序以及无数其他基因组被测定的变革。也是在这 40 年间，DNA 测序技术的应用被不断地创新与扩展，其中就包括用于各种分子水平定量分析。我们认为，纵观整部科学发展史，DNA 测序技术的地位应当不亚于显微镜。

DNA 测序有着两段紧密联系的历史背景，其一是科技的发展，另一个是不断验证其有效性的众多问题。在此，我们首先回顾 DNA 测序技术的发展历史，然后我们来看一下 DNA 测序应用的变化。最后，我们探讨了 DNA 测序的未来。

1 DNA 测序简史

DNA 测序技术的发展不可谓不精彩，在短短数十年间重大突破不断。接下来，我们来回顾在测序技术发展早期，生物学家对生物大分子测序的研究、DNA 测序中电泳技术的发明及其在人类基因组计划中的应用，以及第二代（高通量并行测序）和第三代（单分子实时测序）技术的诞生。Box1 中汇总了一些重要的里程碑事件。

1.1 早期测序

Fred Sanger 认为对于生物大分子的特殊化学结构的了解有助于进一步深入研究，秉持着这个信念，他将他的科学生涯投入到了一级结构的测定之中。没有想到的是，他的工作恰好为此后所有的生物大分子、蛋白、以及 RNA 的测定技术奠定了坚实的基础。

上世纪 50 年代初期，Sanger 测定了第一个蛋白序列，胰岛素蛋白，他将蛋白的两条链裂解，分别测定碎片序列，并通过碎片之间的重叠部分拼出了完整的序列。他的研究明确指出，蛋白质中的氨基酸残基有着特定的排列方式。后来发展出的 Edman 降解法，通过对氨基酸片段末端反复衍生化来测定蛋白序列，简化了蛋白测序。虽然测序的方法很落后，但是在 60 年代晚期的时候，已经测出了很多的蛋白序列，并且证实了相同的蛋白在不同的种属甚至不同个体间都不同。

上世纪 60 年代中，RNA 测序的发展也经历了上述的过程，将一类 RNA 用 RNA 酶

裂解，片段用色谱和电泳分离，每一个片段用不同的核酸外切酶降解测序，最后通过 reads 之间的重叠部分拼成一段完整的序列。丙氨酸 tRNA 是第一个测得序列的 RNA，这项工作令 5 位研究人员花费了 3 年的时间完成，他们使用了 1g 的纯化样品（从 140kg 的酵母中提取）却仅测得了 76bp 的序列。RNA 测序的发展由于指纹图谱技术的出现而大大加速，这项技术将放射性元素标记的 RNA 片段进行二维分离，从而展现出他们的大小和序列信息。

1.2 DNA 测序技术的发明

早期的 DNA 测序技术都十分落后，1968 年，Wu 使用引物扩增技术获得了 lambda 噬菌体的 12bp 粘性末端序列。1973 年，Gilbert 和 Maxam 报导了乳糖阻遏结合位点的 24 个碱基的序列。他们将 DNA 转录成 RNA，并对这些片段进行了测序，这项工作历时两年，平均每月一个碱基。

在 1976 年左右，两种测序技术的出现使得人们可以在一个下午的时间内测得上百个碱基的序列。两种技术分别是：Sanger 与 Coulson 发明的双脱氧链终止法和 Maxam 与 Gilbert 发明的化学裂解法。这两种方法都是将放射性元素标记的片段分离，从而确定核酸序列。Sanger 的方法使用标记的引物通过 DNA 聚合酶进行四轮扩增，每一轮扩增都有特定种类的双脱氧核酸可以使链随机终止，以产生不同长度的 DNA 片段。Gilbert 的方法使用末端标记的 DNA 片段，在四种反应中使用化学试剂产生碱基特异性的裂解片段。这两种方法都使用琼脂糖凝胶电泳来测定每一轮碱基特异性反应中 DAN 片段的大小，这种琼脂糖凝胶可以以单碱基的分辨率分离不同的大小的 DNA 片段。点样时一轮反应点一个孔，使用 X 射线进行放射性自显影，从而使我们可以立即读出条带信息，从下到上依次排列，便是目标序列。

这几项技术一经发布便得到了广泛的应用。在此后的 1979 年，Staden 提出了霰弹枪测序法，即利用对 DNA 片段的随机克隆进行测序，之后又利用片段间的重叠部分拼出完整的 DNA 序列。这项技术由于 Messing 在 1980 年前后对单链 M13 噬菌体克隆载体的改造而加速发展，并被应用到了基因组从头测序。例如在 1982 年发表的 lambda 噬菌体基因组。在 1987 年，Smith、Hood 以及 Applied Biosystem 发明了基于荧光标记的 Sanger 测序仪，可以以每日 1,000 碱基的数据量测序。此后，测序的数据以摩尔定律的速度量疯狂增长，直接促成了中心数据库的建成（如 GenBank）并且通过比对工具（如 BLAST）使这些数据的价值得以体现，也使人们形成了数据共享的共识。在 1982 年时有约 50 万碱基的数据存储在 GenBank 中，而到了 1986 年，有近千万的碱基和全

基因组测序 (Whole Genome Sequencing, WGS)的统计信息存储在 GenBank 中。

1.3 测序技术在 HGP 中的应用

在人类基因组计划 (Human Genome Project, HGP) 中，产生了“多级破碎”(Hierarchical Shotgun) 的概念，在这种方法中，人类基因组的长片段首先被克隆到细菌人工染色体 (Bacterial Artificial Chromosomes, BAC) 中，每一个 BAC 上的 DNA 片段被再次破碎，大小分档，并且再次克隆。对不同的克隆分别进行培养，最后进行 DNA 分离纯化。纯化的 DNA 产物送去进行 Sanger 测序，从凝胶分离的激光扫描图像上获取信号，再进行碱基识别以获得最终的序列。需要注意的是这些实验需要很多步骤，每一个实验都要保证准确性以确保结果的正确性，这便使很多人开始怀疑以这样一种不惜一切代价的方法测得的人类基因组是否正确。

事实上，在大型基因组测序开始之时，对于每一个测序步骤的规模和效率的提升需求日益凸显。这些升级出现在了上世纪 90 年代，关键的升级有：

1) 从荧光标记引物到荧光标记末端，从而使反应在一次完成而不是在分四轮完成；
2) 有了一种对荧光末端标记底物有偏好性的突变型 T7 DNA 聚合酶；
3) 线性扩增反应，减少了对于反应模板量的需求，促进了测序微型化；
4) Oligo (dT) 磁珠的出现简化了测序前处理流程；
5) 双链 DNA 测序方法的出现，使质粒克隆和双端测序变为可能；
6) 毛细管电泳(Capillary Electrophoresis, CE) 代替了凝胶电泳分离，同时也简化了荧光信号的提取和碱基识别；
7) 工业化进程显著增加了测序的效率、减少了测序错误，例如测序自动化、质量控制 (Quality Control, QC)以及操作流程的标准化等。

Wet的实验操作规程只是挑战的一部分，人们在软件的开发上投入了更多精力，例如克隆跟踪，测序数据的碱基识别和拼接等。例如，phred 有着可靠的碱基识别矩阵并且能够区分密切相关的重复性序列，它的出现代替了手工编辑序列的操作。测序得到的 reads 依据重叠部分拼接成为更长的、连续的序列。随着测序目标基因组的复杂性不断提升，重复序列的干扰越来越严重，即使 BAC 测序深度不断提升，但是总有没有连上的 gap，最终导致测序结果的不连续，急需用其他的方法来解决。双端测序的出现，帮助我们将 contigs 连接成有 gap 的 scaffold，从而使我们可以对邻近的 gap 进行直接测序。但也有一些问题只能由人工解决，一些科学家被训练成为“finishers”来评估不同克隆的拼接质量并且将它们连在一起。

从表面看来，技术的发展在稳步推进，但是 90 年代中，并行计算逐渐对于人工决策的替代使得测序的成本显著降低，测序技术得到了快速的发展。2001 年，在一小部分实验室中，已经可以凭借自动化流程每日产生千万碱基的数据量。在 HGP 内外，基因组拼接软件日益成熟，这类工具有：phrap、TIGR 拼接软件和 Celera 拼接软件。这些工具的存在使我们足以应对更加复杂的目标基因组。我们快速增长的测序能力使我们成功测得了几个高质量的基因组，如流感嗜血杆菌(2Mb, 1995)，酿酒酵母(约 12Mb, 1996) 以及秀丽隐杆线虫(约 100Mb，1998). 人类基因组计划的人类基因组，大小约为秀丽隐杆线虫的 30 倍，于 2001 年组装成草图，并于 2004 年最终完成测序。与 HGP 同时期进行的工作还有独自使用全基因组鸟枪测序法进行人类基因组测序的(Whole-Genome Shotgun Sequencing) 的 Craig 和 Celera，此种测序方法也指导了黑腹果蝇的测序(约 175Mb, 2000). 这些测序方法之间的联系在下文会有详细说明。

2004 年，我们已经可以以 1 美元的价格测序 600-700bp，但是基本上已经达到了当时的极限。同时，随着 HGP 的完成，大规模测序的未来将何去何从，无人知晓。

1.4 高通量并行 DNA 测序

在上世纪 80 年代到 90 年代之间，一些研究小组在找寻着除了电泳测序之外的方法。尽管这些努力直到 HGP 完成之后都没有得到回报，但就在他技术成熟后的十年之内，几乎完全取代了 Sanger 测序。大规模并行测序，又叫下一代测序(Next Generation Sequencing, NGS) 。NGS 技术在一些方面与电泳测序技术大相径庭，其关键之处就在于高通量。与一管一个反应不一样，在 NGS 中，建好的 DNA 文库被固定在了二维平面上，这个平面足够每一个模板都进行一次反应。与细菌克隆不一样，每一个模板经过体外扩增然后测序。最终，不进行片段长度测量，测序包含了多组生化反应循环（例如，聚合酶催化的荧光标记核酸的合成）之后成像，也被称作边合成边测序(Sequencing By Synthesis, SBS)。

尽管不一定需要 PCR 的步骤（如单细胞 SBS），但是对于有着成千上万的固化模板的 NGS 技术来说，他的高通量很大程度上是依赖体外扩增技术的。这个过程，又被叫做“香肠式反应”或“桥式 PCR”，使用固定在平面的引物来进行复杂的 DNA 文库扩增，每一个模板都会扩增成紧密的簇状。还有一种方法是 PCR 扩增在试剂中完成，使用 Oligo(dT)磁珠来富集 mRNA，最后排列在测序平面上进行测序。第三种方法是使用扩增来在反应液中产生克隆“纳米球”，之后阵列于平面、测序。

对于 SBS，主要有三种测序方法，第一种是 Ronaghi 与 Nyren 的焦磷酸测序，他们

通过分布逐渐加入每一种 dNTP 来实现。在 dNTP 参与反应的同时会有焦磷酸释放出来，随即激发萤火虫荧光素酶释放的光。与之类似的一种方法是使用离子敏感的晶体管检测 dNTP 的反应。第二种方法是用特殊的 DNA 连接酶将荧光标记的多聚核苷酸与模板序列碱基互补配对连接。第三种方法，也是经久不衰的方法，使用逐渐加入的，荧光标记的脱氧核酸进行 DNA 聚合酶催化的合成反应。SBS 成功的关键因素有三，一是可逆末端终止反应，二是荧光标记的 dNTP，三是经过合理改造的 DNA 聚合酶，从而使得每一个模板在每一轮反应中只结合一个 dNTP。在成像之后，我们可以看到这一轮反应中每一个模板上最新结合的是哪一种 dNTP，随即去除末端的封闭与荧光基团，准备好进行下一轮反应。Solexa 公司（由 Balasubramanian 和 Klenerman 建于 1998 年）便采用了第三种测序方法。

第一个完整的NGS测序平台是在2005年出现的，Shendure, Porreca, Mitra和Church进行了大肠杆菌的重测序；Margulies, Rothberg 和 454 公司进行了生殖支原体的从头测序拼接；Solexa 进行了 phiX174 和人类基因组 BAC 的重测序。这些研究表明，有参考基因组的情况下，短读长的数据也十分有用。在三年之内，Solexa 平台就实现了以读长为 35bp 的双端测序技术进行的人类基因组重测序。

在 2005 年，454 发售了第一款商用 NGS 测序仪。在人类基因组计划之后，大规模基因组测序还只能由几家大型基因组中心完成，而在454及随后发布的几款测序仪的竞争中，个体实验室也可以进行人类基因组时期，大型基因组数据中心才能完成的测序任务。这样“大众化”的测序能力的出现导致了基因组学领域产生了一大批的新技术、新方法、以及其他各个方面不断涌现的创新。

与 ABI 公司在 HGP 时期的一家独大不同，在 NGS 产业的竞争中，有罗氏支持的454平台、Illumina支持的Solexa平台、ABI支持的Agencourt、Quake支持的Helicos、Drmanac 创办的 Complete Genomics 以及 Rothberg 创办的 Ion Torrent，直接导致了在 Florida 的 Marco 岛上举办的 AGBT 会议上各种新型仪器的快速迭代。在 2007-2012 年间，每个 DNA 碱基的测序成本降低了四个数量级。

自 2012 年开始，技术更新的速度开始放缓，然而竞争还在继续。其间，454、SOLiD 以及 Helicos 平台定制了研发，Illumina 平台占据了主导地位（当然，Complete Genome仍然是一个潜在的竞争者）。自2005年以来NGS出现以来，我们获得的成就同样令人惊叹。即便是读长仍然小于 Sanger 测序，但在这 100 多个的碱基中，测序的准确性达到了 99.9%，两台之内，一个刚刚毕业的学生就可以操作测序仪测(Illumina

NovaSeq)出超过十亿的 reads，碱基总量超过 1TB，而这一切只要几千美元。约为 HGP 中所绘制的人类基因组草图的 23G 碱基量的 40 倍之多。

1.5 单分子实时测序技术

几乎我们前面提到的所有测序方式都需要使用模板扩增。然而，扩增会导致复制时出错，出现因序列特点而导致的扩增偏好性和序列信息丢失（如甲基化）的现象，更不用提之后的测序深度和复杂性增加。在理想的情况下，测序应当做到高还原性、高准确度和无限读长。让我们把时间调回到上世纪 80 年代，一部分研究小组为了达到这一目标，采用了比 NGS 更加激进的方法。他们之中有许多研究到最后都不了了之，但是万幸的是有两个保留了下来，也就是我们最近所听到的单分子实时测序技术。

第一种方法，由 Webb 和 Craighead 最先提出，并被 Korlach、Turner 和 Pacific Biosciences (PacBio)进一步改进。这种方法是通过实时检测酶催化的 DNA 聚合反应所释放出的荧光来实现的。其零模波导孔 (zero mode waveguide)的直径小于荧光波长的一半，从而使荧光限制在极小的体系内，之中只有一个聚合酶和相应的模板。因此，只有正在参与 DNA 模板链扩增的荧光标记核酸释放出的荧光信号才足以进行碱基识别。被改造过的聚合酶具有极高的效率，从而使测序数据的读长一般在 10kb 以上，有的甚至可以达到 100kb。PacBio 平台的测序通量比高通量的 NGS 平台如 Illumina 小了一个数量级，但是几年之后两者的差距便不再悬殊。它的错误率大约在 10%，并且随机分布。PacBio 有着偏好性低（如可以兼容高 GC 含量的样品），随机误差分布，超长读长和高测序覆盖度的特点，可以使从头测序中拼接出的序列有着前所未有的准确性和连续性，这对于很多物种来说其重要性不亚于 HGP 对于人类的作用。

第二种方法，是纳米孔测序(nanopore sequencing)，这个概念首先于上世纪 80 年代提出，其主要原理是当一个单链 DNA 通过一个狭长的孔道时，其释放出的离子信号可以反应其一级结构。将这一想法变为现实历经了数十年的研究。首当其冲的就是电场驱动的 DNA 链经过纳米孔时，其通过速度太快以至于释放的信号无法及时捕获。人们对这些问题提出的解决方案有：引入一个名为”ratchet“的酶、对于纳米孔蛋白的鉴定和改造以及对结果信号的更好的检测。这些改进最终使 nanopore 测序在相关产业和学术方面走向成功，主要的经营者是 Bayley 于 2005 年创办的 Oxford Nanopore Technologies (ONT) , 其测序读长于 PacBio 类似，甚至比它更长，而现在所知道的最长读长达到了 900kb。nanopore 测序与之前技术的一个主要区别在于，其无与伦比的便携性。因为他不是依赖于光信号而是使用电信号进行测序的，其体积可以达到 USB 闪存盘大小。虽

然仍有缺陷（如误差可能不是随机分布），但项目正处于积极的开发中。

核酸测序在理想的情况下，应当可以检测到 DNA 修饰现象的存在。在现阶段，PacBio 和 Nanopore 都是可以高度还原共价修饰现象的（如 DNA 甲基化）。单分子的测序方法使我们直接测序完整 RNA 和蛋白质的想法成为可能。

从 1977 年开始，我们历经了 DNA 测序技术的快速革新，并且创新，还在继续。尽管 Illumina 是当前测序届毋庸置疑的巨头，但是之后的市场却不一定是一家独大，并且其他的测序技术很可能会占据重要的低位（例如，PacBio 用于从头测序，Nanopore 用于便携式测序）。NGS 和单分子测序技术都没有在成本和通量上达到顶峰，与此同时也不断有着其他创新的观点涌现，这里便不再赘述（如固态微孔和电镜技术）。不是所有的研究都会得到最终的成果，但上述的内容说明，测序技术的变革需要时间去孕育。

2 DNA 测序技术的应用

DNA 测序技术应用的层面和规模在过去的数十年内不断扩大，其中也有着测序技术发展的推动。下面我们来探讨测序技术应用的几个重要领域，其中包括：基因组从头组装、个体基因组重测序、测序技术的临床应用以及测序仪变为分子计数装置的历程。同样，一些与参考基因组的产生、技术应用和相关软件的发展相关的重要里程碑事件，已经在 Box2 中列出。

2.1 基因组从头测序的拼接

对于 DNA 测序的头 25 年来说，他的主要目的就是为了测出部分或是完整的基因组。事实上，在 Sanger 测序技术提出的同时，也发表了第一个基因组(phiX174; 5.4kb)，而这个基因组主要是由手工拼接的。DNA 测序只是为数不多的几项能够应用拼接技术的研究。如果 DNA 序列是随机的，那么我们仅仅依据重叠片段拼出的大型基因组会有无数种可能，万幸的是，他不是随机的。但是，系统误差和重复性序列共同注定了我们不可能仅凭上千个剪辑就拼接得到高质量的大型基因组。我们还需要一些能够把他们连接的“连续性信息”。

在 HGP 中，这些附加的信息包括：

1) 遗传图谱，是根据同一家系中，遗传多样性造成的生殖隔离绘制的。它可以在染色体水平上提供序列次序的定位信息。
2) 所克隆 BAC 的物理图谱，通过限制性内切酶的指纹图谱来确定重叠区域和回贴基因组时的位置。基因的克隆是被分别打断并测序的，所以依据物理图谱将不同的重复序列分开，依次拼接便可得到完整信息。

3) 双端测序，由 Ansorge 在 1990 年提出，结果由大致确定长度的 DNA 片段两头测序的 reads 组成，可以有效地将这些末端序列连接起来。因为克隆方法的不同，片段长度可能从几千 bp 到数十万 bp 不等。通过 8-10 倍的测序深度，与上述的连续性信息结合，不仅可以进行基因组拼接，还可以将大多数基因组的错误率降低至十万分之一碱基。

除上述方法之外，人们还进行了其他的实验来填补拼接的 contigs 之间的 gap 或是降低误差。Celera 一直致力于双端测序的发展，并且尽量避免物理图谱的辅助。其中重要的进展是应用贪婪算法软件的出现，如 phrap 和 TIGR，和 Celera 的拼接软件的，以图形为基础的拼接方法（重叠片段-铺展-取一致性序列）。尽管 Celera 的拼接方法有着一定的合理性，但是由于大量重复序列的干扰，其算法正如 HGP 时期的克隆方法一样，不能仅凭自身组装出高质量基因组。现在的人类参考基因组由参考基因组组织(Genome Reference Consortium)继续维护，会不时发布一些参考基因组的更新。

NGS 从 2005 年发展至今，从头测序的基因组组装得到了长足的发展。从表面上看来，短读长和重复序列多的问题因新算法（如基于 de Brujin graph 的 EULER 与 Velvet）的出现而得以解决，但实际上，当我们将其应用到更大的基因组上时，与 HGP 的参考基因组相比，其组装质量还是有待提高。尽管短读长是一部分原因，但我们不能把目光局限于此。事实上，真正导致组装质量低的原因是缺乏一套完整的流程。双端测序可能会起到一定的效果，但是体外建库的方法可能会对那些本可以跨越的距离要求更加苛刻。另外，这个领域还缺乏遗传和物理图谱层面上的大规模并行处理方法。

这样的低谷并未持续多久，我们有充足的理由对从头测序的未来充满信心，首先，与多级鸟枪测序法相似的，对于高分子量的基因组片段进行体外建库的方法已然建立。其次，一些如 Hi-C(全基因组染色体结构捕获 genome-wide chromosome conformation capture) 和光学图谱的方法，可以提供一种行之有效的、将组装在 scaffold 水平的基因组定位到染色体水平的方法。再次，PacBio 和 ONT 测序的读长已经可达到数十万碱基，现在的主要限制是在于对于长基因组片段体外建库的方法而不是测序技术本身。PacBio 成功测序了高 GC 含量的细菌基因组，从而证明了在单分子测序中免去克隆和扩增的优势。为了解决长读长带来的高错误率和多平台数据混和拼接的问题，人们再一次开始使用 Celera 软件曾使用的拼接方法。通过把长读长和更广泛的连续性信息结合起来（如长片段基因组文库建库、核染色质联合分析与光学图谱等方法），让我们看到了使用“后 Sanger 测序法”从头测序组装出无损人类参考基因组的希望。

2.2 基因组重测序技术

在 HGP 之后，下一个目标便是整理人类的遗传变异，便是我们所说的重测序，因为 Sanger 测序的价格居高不下，重测序被用来发掘大范围的差异，也就是后来发展出的基因芯片，其性价比极高，大大加速了广谱基因组学的联合分析研究进展。为了改变这一现象，人们提出了 1000 美元基因组计划，这样的价格几乎是拼接第一个人类基因组所花费价格的百万分之一。这一计划最早在 2001 年提出（于加利福尼亚大学，Santa Cruz 人类基因研讨会），那时 NGS 还没有出现，并在几年后，由国立人类基因组研究所(National Human Genome Research Institute, NHGRI) 的创新 DNA 测序技术项目进一步完善。项目提供了 2.2 亿美元的资金，用以支持 40 个学术团体，还有 27 个来源于各大商业平台的经济实体，对项目给予了各种直接或间接地帮助，在很大程度上促进了上述进程的发展。

重测序，不同于基因组拼接，是一种将序列的测序数据回贴到基因组上，以鉴定遗传差异的测序方法。一些新的算法，比如 Bowtie 和 Burrows-Wheeler Aligner (BWA)，借用了数据压缩技术的理念，将数百万条序列高效的回贴到基因组上。人们采取高测序深度（如 30X）来鉴定杂合序列，同时也尽量将真实的遗传差异和测序错误区分开来。一些流行的软件包，比如 SAMtools 和之后的 GATK，接替了 phred，开始用作处理 NGS 碱基、数据和变异的分析工具。短读长的数据，特别是在双端测序的情况下，可以被特异性回贴到大部分的参考基因组上。但很多情况下不会完全回帖上去，并且短读长重测序的另一个缺点是，重复性序列区域和结构变异区域的遗传差异常常会丢失。而克服这一缺陷的方法是使用 PacBio 来对人类基因组进行重测序，第二个造成回帖不完整的原因是二倍体植物所处的不同时期，也就是单倍体所造成的影响。所幸单倍体造成的问题可以被一些与从头测序拼接串联方法类似的方式解决（并且在理想的情况下，从头测序是不会受到单倍体的干扰的）。尽管这些方法没有被广泛使用，但是这些方法的可行性正不断提升。

HGP 的基因组是基于不同的志愿者贡献的 DNA 构建的，但是大部分的数据是由一个人得来的，这个人就是纽约的 Buffalo，他认为一部分欧洲人是非洲的祖先迁移过来的后代。Craig Venter 在 2007 年第一次接受了全基因组重测序，也是 Celera 基因组中的一个样本，有着许多附加数据的补充。紧接着，第二个接受测序的 Jim Watson 在 2008 年于 454 平台上接受了测序，此后又有两位匿名志愿者和一位患者的生殖细胞和肿瘤基因组于 Illumina 的 Solexa 平台上接受了检测，五位志愿者在 Complete Genome 平台接受

检测。在这个时期内，全基因组测序对于大部分组织来说还太过昂贵，于是促进了特定位点测序方法的出现，之后又出现了全外显子测序技术，即对基因组中编码蛋白序列的 1-2%的序列进行测序。

随着 WGS 成本控制在 1000 美元，而 WES 测序只要几百美元，人类个体重测序的进程大大加速。2008 年提出的 1000 基因组计划，在 2010 年公布了几百个低覆盖度的 WGS 结果，到了 2015 年，这个数字变为数千个。外显子组测序项目在 2013 年公布了超出 6500 个外显子组数据。最近建成的基因组整合数据库(Genome Aggregation Database, http://gnomad.broadinstitute.org/) 整合了超过 120,000 个外显子组数据和超过 15,000 个基因组。Genomics England (https://www.genomicsengland.co.uk/), GenomeAsia100K (http://www.genomeasia100k.com/) 和 NHLBI TOPMed (TransOmics for Precision Medicine, https://www.nhlbiwgs.org/) 项目，都计划在未来的一到两年内完成超过 100,000 个个体的测序任务。这些计划仅仅是测序界项目的冰山一角，而 WEG 与 WGS 共计已完成超过一百万人类的重测序，值得庆贺。

2.3 测序技术的临床应用

我们对于人类基因组的测定在很大程度上促进了遗传变异的阐释，也在一定程度上说明了为什么临床用药还很大程度上依赖于 WGS。DNA 测序已经在数个领域中证明其有效性，我们在此列出其中的三个

DNA 测序在临床应用中最有代表性的是无创产前检查(non-invasive prenatal testing, NIPT). Lo 与 Quake 于 2008 年的研究表明，对胎儿释放在血液中的 DNA 片段进行检测，可以检查出其染色体的非整倍现象。基于测序技术的产前筛查比历史上所有的分子检测技术都要快，并且世界上已经有数百万的产妇从这种为 NIPT 而量身打造的低通量测序中受益。

WES 的早期应用之一是快速发现患者因孟德尔遗传疾病而发生改变的新基因，以用于诊断这类遗传疾病。之后人们又很快发现神经性发育失调疾病的发生很大程度上是由于编码区域的原发性突变所导致的。慢慢地，WES 成为诊断孟德尔遗传疾病和神经性发育失调疾病的重要工具之一，在所有诊断出孟德尔疾病的患者中，有 25%是通过 WES 确诊的，并且这个数字还在不断升高。

在我们对于癌症的认识中，它是一种基因层面的疾病，而这种认识正在被 DNA 测序所扭转，大规模重测序暴露了癌症的遗传不稳定性，也很好地规范了其分子鉴定技术。DNA 测序技术对癌症的临床诊疗的影响有：

1) 有助于对存在于不同个体的癌症突变进行靶向治疗；

2) 利用癌组织周围的游离细胞或体液中游离的 DNA 进行无创检查或监测；

3) 鉴定癌症特有的，可以调控蛋白的突变可能会有助于个人疫苗的制备。

尽管在这些领域成功的先例还很少，对于癌症的治疗效果目前也有限，但是研究仍在进行中。

2.4 测序仪，分子定量装置

1983 年时，外显子标签被认为是一条发现新基因的捷径，SAGE(Serial Analysis of Gene Expression)技术引入了使用测序方法来对基因的表达量进行数字定量。SAGE 是将从 cDNA 上扩增得到的序列加上了足够回贴到原始基因上的标签，串联起来进行 Sanger 测序的方法。早在 2000 年的时候，Brenner 和 Lynx 公司提出了对 cDNA 标签进行“大规模并行测序检测”的设想，是 NGS 的前身。然而，直到 2008 年，五个研究小组进行了 RNA-Seq 的时候，这一理念才被广泛接受。RNA-Seq 使用 NGS 平台，对 cDNA 的 3‘段片段或全长进行鸟枪测序，从而对转录组进行定量分析与鉴定。RNA-Seq 在基因芯片方面的应用有着很大的优势，其中最引人注目的是，转录组定量带来的对于统计方法的更新，也有许多软件包助力这一进程，包括 TopHat 与 Cufflinks。

也是在 2008 年左右，一些小型实验室开始使用 NGS 设备开始了数字化定量的研究，领域涉及转录因子结合位点，染色质鉴定及翻译等。在接下来的数十年内，DNA 测序技术作为一种分子定量装置，在大量的生物医学领域和分子现象研究中发展出了上百种实验方法，这些应用领域包括：转录、翻译、DNA 复制、核酸修饰、转录后修饰、核酸蛋白互作、蛋白互作研究等。这些研究在其他的综述和资料之中多有整理（http://enseqlopedia.com/），便不再赘述。

测序仪作为分子定量装置的应用很快便得到了广泛的认可，而且也许在推动 NGS 在生物医学应用方面有着比拼接和重测序更为重要的作用。DNA 测序仪之于分子生物学家，恰似显微镜之于细胞生物学家，都是一种重要而又基础的测量工具。从长远来看，这个比喻对于 DNA 测序技术来说再恰当不过了。

2.5 宏基因组测序

即使用鸟枪测序法对复杂的微生物群落测序。例如，对于环境的宏基因组和人体微生物菌群进行测序，这就带来了其研究的独特性，需要将拼接、重测序、和定量分析结合在一起。一些近期的综述已经对这一领域的进展进行了总结。

3 DNA 测序的未来

纵观整部科技发展史，DNA 测序技术还很年轻，这里，我们从已有的领域和正在分化的研究方向中简单的预测一下 DNA 测序技术的未来。

3.1 基因组差异分析

现在，仅有一个真核细胞完成了真正意义上的完整基因组测序，即对每一个染色体从一个端粒到另一端的数据中没有空缺也没有歧义。随着测序技术的发展，我们有理由相信我们可以挑战一些附加基因组的测定，例如着丝粒。地球上有着成千上万的物种（灭绝了的物种会更多），每一个物种都有一个基因组等着我们去测定，还有这难以计数的微生物组和宏基因组可以去测。我们想，在一些设想不到的方面，全面分析所有的基因组会相当有效，例如蛋白结构测定。

3.2 种群水平重测序

我们已经有约 0.1%的现存人类对他们的基因组进行了一定水平的重测序，这是一个具有里程碑意义的进展。对于我们祖先和一些古人类的重测序正在刷新着我们对于人类历史的了解。在近几代出现的原始变异位点数量远远超过了人类基因组中核算的数量。在未来，难以计数的基因组或许会让我们可以绘制出人类基因组的进化历程（即观察所有与生命有关的杂合基因变异现象）。DNA 测序在未来也对法医学有着重大意义，可以使我们不用从鉴定对象上获取样品而鉴定其身份。

3.3 发育生物学

我们每一个人都是从单细胞发育而来的。然而，我们对于发育的了解还很少。当今的科技发展已经允许我们对单细胞进行基于测序技术的分析。尽管常用的方法还需要在体外完成（如单细胞转录组测序），即使还收到周围空间环境的限制，但已经有更为先进的方法可以对 RNA 或蛋白质进行原位测序。其他的策略有使用体外基因组编辑技术来追踪细胞间遗传关系，或者转运标记物来明确神经网络的连接关系。对于 DNA 的编辑有可能使我们在更宏观的层面上记录生物学现象，例如监测基因表达或钙离子含量。

3.4 实时、便携检测

Nanopore 测序仪重量仅有 70g，可以在上样 30 十分钟内开始产出数据。试想，如果有着大量的 Nanopore 测序仪，我们就可以全面监控我们的核算变化、环境情况、以及我们的日常起居，例如可以深入追踪我们的空气、食物和身体情况。大量装置所产生的数据还可以通过我们 GPS 和音视频系统整合到一起。

3.5 还有更多

DNA 测序技术还可能在一些意想不到的地方大展身手，例如，NGS 用作在合成 DNA 上读取预存储的数据。Nanopore 也可以在测序领域之外另辟蹊径，例如监控待测物结合情况、化学纳米计算机、以及蛋白的折叠和解构。

4 DNA 测序，新的“显微镜”

距离光学显微镜的发明已有 400 余年，而这项技术还会被继续使用和发展下去。而反观 DNA 测序仅有短短 40 余年，这项技术在未来的数十年甚至几个世纪中也将继续发展。就其改变的生物医学研究的速度，以及其开始改变临床医学的趋势来讲，我们预测，DNA 测序技术必将长久存在，其的影响将不亚于显微镜，甚至影响更为深远。

译文原文

REVIEW

doi:10.1038/nature24286

DNA sequencing at 40: past, present and future

Jay Shendure[1,2], Shankar Balasubramanian[3,4], George M. Church[5], Walter Gilbert[6], Jane Rogers[7], Jeffery A. Schloss[8] & Robert H. Waterston[1]

This review commemorates the 40th anniversary of DNA sequencing, a period in which we have already witnessed multiple technological revolutions and a growth in scale from a few kilobases to the first human genome, and now to millions of human and a myriad of other genomes. DNA sequencing has been extensively and creatively repurposed, including as a 'counter' for a vast range of molecular phenomena. We predict that in the long view of history, the impact of DNA sequencing will be on a par with that of the microscope.

DNA sequencing has two intertwined histories—that of the underlying technologies and that of the breadth of problems for which it has proven useful. Here we first review major developments in the history of DNA sequencing technologies (Fig. 1). Next we consider the trajectory of DNA sequencing applications (Fig. 2). Finally, we discuss the future of DNA sequencing.

History of DNA sequencing technologies

The development of DNA sequencing technologies has a rich history, with multiple paradigm shifts occurring within a few decades. Below, we review early efforts to sequence biopolymers, the invention of electrophoretic methods for DNA sequencing and their scaling to the Human Genome Project, and the emergence of second (massively parallel) and third (real-time, single-molecule) generation DNA sequencing. Some key technical milestones are also summarized in Box 1.

Early sequencing

Fred Sanger devoted his scientific life to the determination of primary sequence, believing that knowledge of the specific chemical structure of biological molecules was necessary for a deeper understanding[1]. Ironically, given the state of sequencing technology for each biopolymer today, proteins and RNA came first.

The first protein sequence, of insulin, was determined in the early 1950s by Sanger, who fragmented its two chains, deciphered each fragment and overlapped the fragments to yield a complete sequence. His work showed unequivocally that proteins had defined patterns of amino acid residues[2]. The later development of Edman degradation, a repeated elimination of an N-terminal residue from the peptide chain, made protein sequencing easier[3]. Although these methods were cumbersome, many proteins had been sequenced by the late 1960s, and it became clear that each protein's sequence varied across species and between individuals.

In the 1960s, RNA sequencing was tackled by this same general process: an RNA species was first fragmented with RNases, next the pieces were separated by chromatography and electrophoresis, then individual fragments were deciphered by sequential exonuclease digestion, and finally the sequence was deduced from the overlaps. The first RNA sequence, of alanine tRNA, required five people working three years with one gram of pure material (isolated from 140 kg of yeast) to determine 76 nucleotides[4]. This process was greatly simplified by 'fingerprinting' techniques, which included the separation of radioactively labelled RNA fragments and visualization in two dimensions, with the resulting positions diagnostic of their size and sequence[5].

The invention of DNA sequencing

Early attempts to sequence DNA were cumbersome. In 1968, Wu reported the use of primer extension methods to determine 12 bases of the cohesive ends of bacteriophage lambda[6]. In 1973, Gilbert and Maxam reported 24 bases of the lactose-repressor binding site, by copying it into RNA and sequencing those fragments. This took two years: one base per month[7].

The development, in around 1976, of two methods that could decode hundreds of bases in an afternoon transformed the field. Both methods—the chain terminator procedure developed by Sanger and Coulson, and the chemical cleavage procedure developed by Maxam and Gilbert—used distances along a DNA molecule from a radioactive label to positions occupied by each base to determine nucleotide order. Sanger's method involved four extensions of a labelled primer by DNA polymerase, each with trace amounts of one chain-terminating nucleotide, to produce fragments of different lengths[8]. Gilbert's method took a terminally labelled DNA-restriction fragment, and, in four reactions, used chemicals to create base-specific partial cleavages[9]. For both methods, the sizes of fragments present in each base-specific reaction were measured by electrophoresis on polyacrylamide slab gels[10], which enabled separation of the DNA fragments by size with single-base resolution. The gels, with one lane per base, were put onto X-ray film, producing a ladder image from which the sequence could be read off immediately, going up the four lanes by size to infer the order of bases.

These methods came into immediate use. Shotgun sequencing—sequencing of random clones followed by sequence assembly based on the overlaps—was suggested by Staden in 1979[11], greatly facilitated by Messing's development of the single-stranded M13 phage cloning vector around 1980[12], and used to assemble genomes *de novo*, such as bacteriophage lambda as early as 1982[13]. By 1987, automated, fluorescence-based Sanger sequencing machines, developed by Smith, Hood and Applied Biosystems[14,15], could generate around 1,000 bases per day. Sequence data grew exponentially, approximating Moore's law and motivating the creation of central data repositories (such as GenBank) that, through search tools (such as BLAST[16]), amplified the value of each sequence and engendered a spirit of data sharing. By 1982, over half a million bases had been deposited in GenBank; by 1986, nearly 10 million bases (GenBank and WGS Statistics; https://www.ncbi.nlm.nih.gov/genbank/statistics/).

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. [3]Department of Chemistry, University of Cambridge, Cambridge, UK. [4]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. [5]The Wyss Institute & Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge Massachusetts, USA. [7]International Wheat Genome Sequencing Consortium, Little Eversden, Cambridge, UK. [8]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

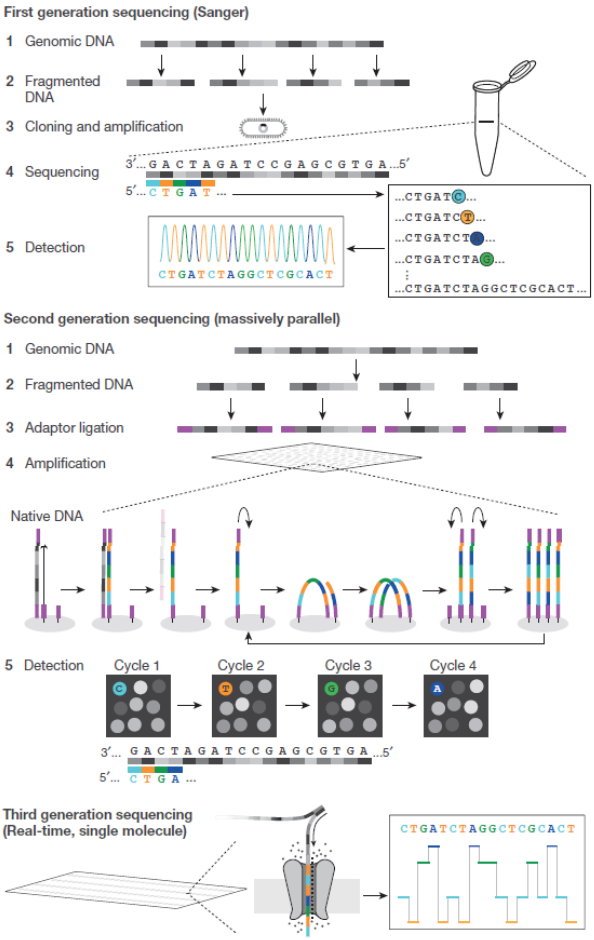

Figure 1 | DNA sequencing technologies. Schematic examples of first, second and third generation sequencing are shown. Second generation sequencing is also referred to as next-generation sequencing (NGS) in the text.

Scaling to the human genome

For the 'hierarchical shotgun' strategy that emerged as the workhorse of the Human Genome Project (HGP), large fragments of the human genome were cloned into bacterial artificial chromosomes (BACs). DNA from each BAC was fragmented, size-selected and sub-cloned. Individual clones were picked and grown, and the DNA was isolated. The purified DNA was used as a template for automated Sanger sequencing, the signal was extracted from laser-scanned images of the gels, and bases were called to finally produce the sequence. The fact that this process involved many independent steps, each of which had to work well, led sceptics to doubt it could ever be made efficient enough to sequence the human genome at any reasonable cost.

Indeed, as efforts to sequence larger genomes took shape, it became clear that the scale and efficiency of each step needed to be vastly increased. This was achieved in fits and spurts in the 1990s. Noteworthy improvements included: (1) a switch from dye-labelled primers to dye-labelled terminators, which allowed one rather than four sequencing reactions[17]; (2) a mutant T7 DNA polymerase that more readily incorporated dye-labelled terminators[18]; (3) linear amplification reactions, which greatly reduced template requirements and facilitated miniaturization[19]; (4) a magnetic bead-based DNA purification method that simplified automation of presequencing steps[20]; (5) methods enabling sequencing of double-stranded DNA, which enabled the use of plasmid clones and therefore paired-end sequencing; (6) capillary electrophoresis, which eliminated the pouring and loading of gels, while also simplifying the extraction and interpretation of the fluorescent signal[21]; (7) adoption of industrial processes to maximize efficiencies and minimize errors (for example, automation, quality control, standard operating procedures, and so on).

Wet laboratory protocols were only half the challenge. Substantial effort was invested into the development of software to track clones, and into the interpretation and assembly of sequence data. For example, manual editing of the sequence reads was replaced by the development of phred, which introduced reliable quality metrics for base calls and helped sort out closely related repeat sequences[22]. Individual reads were then assembled from overlaps in a quality-aware fashion to generate long, continuous stretches of sequence. As more complex genomes were tackled, repetitive sequences were increasingly confounding. Even after deep shotgun sequencing of a BAC, some sequences were not represented, resulting in discontinuities that had to be tackled with other methods. Paired-end sequencing[23] helped to link contigs into gapped scaffolds that could be followed up by directed sequencing to close gaps. Some problems were only resolved by eye; scientists who were trained 'finishers' assessed the quality and signed off on the assembled sequence of individual clones[22].

Although the process remained stable in its outlines, rapid-fire improvements led to steady declines in the cost throughout the 1990s, while parallel advances in computing increasingly replaced human decision making. By 2001, a small number of academic genome centres were operating automated production lines generating up to 10 million bases per day. Software for genome assembly matured both inside and outside of the HGP, with tools, such as phrap, the TIGR assembler and the Celera assembler, able to handle genomes of increasing complexity[22,24,25]. A yearly doubling in capacity enabled the successful completion of high-quality genomes beginning with *Haemophilus influenza* (around 2 Mb, 1995) followed quickly by *Saccharomyces cerevisiae* (around 12 Mb, 1996) and *Caenorhabditis elegans* (around 100 Mb, 1998)[26–28]. The HGP's human genome, which is 30 times the size of *C. elegans* and with much more repetitive content, came first as a draft (2001) and then as a finished sequence (2004)[29,30]. The HGP was paralleled by a private effort to sequence a human genome by Craig Venter and Celera (2001)[31] with a whole-genome shotgun strategy piloted on *Drosophila melanogaster* (around 175 Mb; 2000)[32]. The strategic contrasts between these projects are further discussed below.

By 2004, instruments were churning out 600–700 bp at a cost of US$1 per read, but creating additional improvements was an increasingly marginal exercise. Furthermore, with the completion of the HGP, the future of large-scale DNA sequencing was unclear.

Massively parallel DNA sequencing

Throughout the 1980s and 1990s, several groups explored alternatives to electrophoretic sequencing. Although these efforts did not pay off until after the HGP, within a decade of its completion, 'massively parallel' or 'next-generation' DNA sequencing (NGS) almost completely superseded Sanger sequencing. NGS technologies sharply depart from electrophoretic sequencing in several ways, but the key change is multiplexing. Instead of one tube per reaction, a complex library of DNA templates is densely immobilized onto a two-dimensional surface, with all templates accessible to a single reagent volume. Rather than bacterial cloning, *in vitro* amplification generates copies of each template to be sequenced. Finally, instead of measuring fragment lengths, sequencing comprises cycles of biochemistry (for example, polymerase-mediated incorporation of fluorescently labelled nucleotides) and imaging, also known as 'sequencing-by-synthesis' (SBS).

Although amplification is not strictly necessary (for example, single-molecule SBS[33–35]), the dense multiplexing of NGS, with millions to billions of immobilized templates, was largely enabled by clonal *in vitro* amplification. The simplest approach, termed 'polonies' or 'bridge amplification', involves amplifying a complex template library with primers immobilized on a surface, such that copies of each template remain tightly clustered[36–39]. Alternatively, clonal PCR can be performed in an emulsion,

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

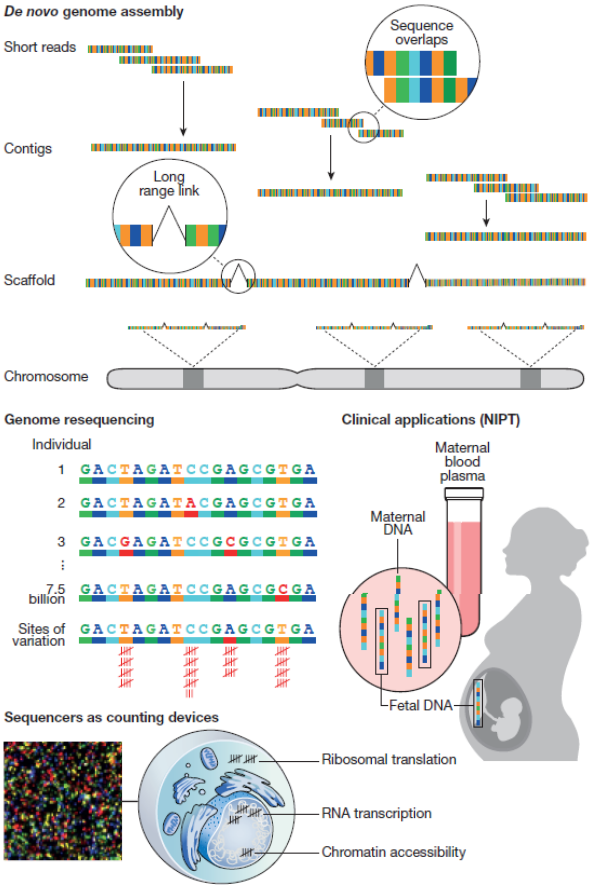

Figure 2 | DNA sequencing applications. Major categories of the application of DNA sequencing include *de novo* genome assembly, individual genome resequencing, clinical applications such as non-invasive prenatal testing, and using sequencers as counting devices for a broad range of biochemical or molecular phenomena.

such that copies of each template are immobilized on beads that are then arrayed on a surface for sequencing[40–42]. A third approach involves rolling circle amplification in solution to generate clonal 'nanoballs' that are arrayed and sequenced[43].

For SBS, there were three main strategies. The pyrosequencing approach of Ronaghi and Nyrèn involves discrete, step-wise addition of each deoxynucleotide (dNTP). Incorporation of dNTPs releases pyrophosphate, which powers the generation of light by firefly luciferase[44]. With an analogous approach, natural dNTP incorporations can be detected with an ion-sensitive field effect transistor[45,46]. A second strategy uses the specificity of DNA ligases to attach fluorescent oligonucleotides to templates in a sequence-dependent manner[41,43,47,48]. A third approach, which has proven the most durable, involves the stepwise, polymerase-mediated incorporation of fluorescently labelled deoxynucleotides[33,34,49]. Critical to the success of polymerase-mediated SBS, was the development of reversibly terminating, reversibly fluorescent dNTPs, and a suitably engineered polymerase[50], such that each template incorporates one and only one dNTP on each cycle. After imaging to determine which of four colours was incorporated by each template on the surface, both blocking and fluorescent groups are removed to set up the next extension[51–53]; this general approach was used by Solexa, founded by Balasubramanian and Klenerman in 1998.

The first integrated NGS platforms came in 2005, with resequencing of *Escherichia coli* by Shendure, Porreca, Mitra and Church[41], *de novo* assembly of *Mycoplasma genitalium* by Margulies, Rothberg and 454 (ref. 40), and resequencing of phiX174 and a human BAC by Solexa[54]. These studies demonstrated how useful even very short reads are, given a reference genome to which to map them. Within three years, human genome resequencing would become practical on the Solexa platform with 35-bp paired reads[55].

In 2005, 454 released the first commercial NGS instrument. In the wake of the HGP, large-scale sequencing was still the provenance of a few genome centres. With 454 and other competing instruments that followed closely after, individual laboratories could instantly access the capacity of an entire HGP-era genome centre. This 'democratization' of sequencing capacity had a profound impact on the culture and composition of the genomics field, with new methods, results, genomes and other innovations arising from all corners.

In contrast to the monopoly of Applied Biosystems during the HGP, several companies, including 454 (acquired by Roche), Solexa (acquired by Illumina), Agencourt[47,48] (acquired by Applied Biosystems), Helicos[34,35] (founded by Quake), Complete Genomics[43] (founded by Drmanac) and Ion Torrent[46] (founded by Rothberg), intensely competed on NGS, resulting in a rapidly changing landscape with new instruments that were flashily introduced at the annual Advances in Genome Biology and Technology (AGBT) meeting in Marco Island, Florida. Between 2007 and 2012, the raw, per-base cost of DNA sequencing plummeted by four orders of magnitude[56].

Since 2012, the pace of improvement has slowed, as has the competition. The 454, SOLiD and Helicos platforms are no longer being developed, and the Illumina platform is dominant (although Complete Genomics[43] remains a potential competitor). Nonetheless, it is astonishing to consider how far we have come since the inception of NGS in 2005. Read lengths, although still shorter than Sanger sequencing, are in the low hundreds of bases, and mostly over 99.9% accurate. Over a billion independent reads, totalling a terabase of sequence, can be generated in two days by one graduate student on one instrument (Illumina NovaSeq) for a few thousand dollars. This exceeds the approximately 23 gigabases that were generated for the HGP's draft human genome by a factor of 40.

Real-time, single-molecule sequencing

Nearly all of the aforementioned platforms require template amplification. However, the downsides of amplification include copying errors, sequence-dependent biases and information loss (for example, methylation), not to mention added time and complexity. In an ideal world, sequencing would be native, accurate and without read-length limitations. To reach this goal, stretching back to the 1980s, a handful of groups explored even more radical approaches than NGS. Many of these were dead ends, but at least two approaches were not, as these have recently given rise to real-time, single-molecule sequencing platforms that threaten to upend the field once again.

A first approach, initiated by Webb and Craighead and further developed by Korlach, Turner and Pacific Biosciences (PacBio), is to optically observe polymerase-mediated synthesis in real time[57,58]. A zero mode waveguide, a hole less than half the wavelength of light, limits fluorescent excitation to a tiny volume within which a single polymerase and its template reside. Therefore, only fluorescently labelled nucleotides incorporated into the growing DNA chain emit signals of sufficient duration to be 'called'. The engineered polymerase is highly processive; reads over 10 kb are typical, with some reads approaching 100 kb. The throughput of PacBio is still over an order of magnitude less than the highest-throughput NGS platforms, such as Illumina, but not so far from where NGS platforms were a few years ago. Error rates are very high (around 10%) but randomly distributed. PacBio's combination of minimal bias (for example, tolerance of extreme GC content), random errors, long reads and redundant coverage can result in *de novo* assemblies of unparalleled quality with respect to accuracy and contiguity, for many species exceeding what would be possible even with efforts similar to the HGP.

A second approach is nanopore sequencing. This concept, which was first hypothesized in the 1980s[59–61], is based on the idea that patterns in

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

BOX 1

The milestones listed below correspond to key developments in the evolution of sequencing technologies. This is a large topic, and we apologize for any omissions.

Technical milestones

1953: Sequencing of insulin protein[2]
1965: Sequencing of alanine tRNA[4]
1968: Sequencing of cohesive ends of phage lambda DNA[6]
1977: Maxam–Gilbert sequencing[9]
1977: Sanger sequencing[8]
1981: Messing's M13 phage vector[12]
1986–1987: Fluorescent detection in electrophoretic sequencing[14,15,17]
1987: Sequenase[18]
1988: Early example of sequencing by stepwise dNTP incorporation[139]
1990: Paired-end sequencing[23]
1992: Bodipy dyes[140]
1993: *In vitro* RNA colonies[37]
1996: Pyrosequencing[44]
1999: *In vitro* DNA colonies in gels[38]
2000: Massively parallel signature sequencing by ligation[47]
2003: Emulsion PCR to generate *in vitro* DNA colonies on beads[42]
2003: Single-molecule massively parallel sequencing-by-synthesis[33,34]
2003: Zero-mode waveguides for single-molecule analysis[57]
2003: Sequencing by synthesis of *in vitro* DNA colonies in gels[49]
2005: Four-colour reversible terminators[51–53]
2005: Sequencing by ligation of *in vitro* DNA colonies on beads[41]
2007: Large-scale targeted sequence capture[93–96]
2010: Direct detection of DNA methylation during single-molecule sequencing[65]
2010: Single-base resolution electron tunnelling through a solid-state detector[141]
2011: Semiconductor sequencing by proton detection[142]
2012: Reduction to practice of nanopore sequencing[143,144]
2012: Single-stranded library preparation method for ancient DNA[145]

the flow of ions, which occur when a single-stranded DNA molecule passes through a narrow channel, will reveal the primary sequence of the strand. Decades of work were required to go from concept to reality. Firstly, electric field-driven transport of DNA through a nanometre-scale pore is so fast that the number of ions per nucleotide is insufficient to yield an adequate signal. Solutions have eventually been developed to these and other challenges, including interposing an enzyme as a 'ratchet', identifying and engineering improved nucleopore proteins, and better analytics of the resulting signals[62]. These advances recently culminated in successful nanopore sequencing, in both academia[63] and industry, most prominently by Oxford Nanopore Technologies (ONT), founded by Bayley in 2005. Sequence read lengths of ONT are on par with or exceed the reads generated by PacBio; with the longest obtained reads presently at 900 kilobases (ref. 64). A major differentiator from other sequencing technologies is the extreme portability of nanopore devices, which can be as small as a memory (USB) stick, because they rely on the detection of electronic, rather than optical, signals. Important challenges remain (for example, errors may not be randomly distributed), but progress is rapid.

Nucleic-acid sequencing would ideally also capture DNA modifications. Indeed, both PacBio and nanopore sequencing have demonstrated the detection of native covalent modifications, such as methylation[64,65]. Single-molecule methods also open up the intriguing possibility of directly sequencing RNA[66,67] or even proteins[68–71].

Since 1977, DNA-sequencing technology has evolved at a fast pace and the landscape continues to change shift under our feet. Although Illumina is presently the dominant supplier of sequencing instruments, the commercial market is no longer monolithic and other technologies may successfully occupy important niches (for example, PacBio for *de novo* assembly and ONT for portable sequencing). Neither NGS nor single-molecule methods have fully plateaued in cost and throughput, and there are additional concepts that are still in development, which are not discussed here (for example, solid-state pores and electron microscopy)[70,71]. Not all will work out, but as the above examples make clear, transformative sequencing technologies can take decades to mature.

Applications of DNA sequencing

The range and scope of DNA sequencing applications has also expanded over the past few decades, shaped in part by the evolving constraints of sequencing technologies. Below we review key areas of application including *de novo* genome assembly, individual genome resequencing, sequencing in the clinic and the transformation of sequencers into molecular counting devices. Some key milestones for the generation of reference genomes and development of applications and software are summarized in Box 2.

De novo genome assembly

For its first 25 years, the primary purpose of DNA sequencing was the partial or complete sequencing of genomes. Indeed, the inception of Sanger sequencing in 1977 included the first genome (phiX174; 5.4 kb), essentially assembled by hand[72]. However, DNA sequencing was only one of several technologies that enabled assembly of larger genomes. If the DNA sequence was random, arbitrarily large genomes could be assembled to completion solely based on fragment overlaps. However, it is not random, and the combination of repetitive sequences and technical biases makes it impossible to obtain high-quality assemblies of large genomes from kilobase-scale reads alone. Additional 'contiguity information' is required.

For the HGP[29,30], these additional sources of contiguity information included the following. (1) Genetic maps, which were based on the segregation of genetic polymorphisms through pedigrees, that provided orthogonal information about the order of sequences locally and at the scale of chromosomes. (2) Physical maps, for which BACs were cloned, restriction-enzyme 'fingerprinted' to identify overlaps and ordered into a 'tiling path' that spanned the genome. Clones were individually shotgun sequenced and assembled, thereby isolating different repeat copies from one another, and then further ordered and assembled. (3) Paired-end sequencing, introduced by Ansorge in 1990[23], comprises sequencing into both ends of a DNA fragment of approximately known length, effectively linking those end-sequences. Depending on the cloning method, the spanned length could range from a few kilobases to a few hundred kilobases. Sequence coverage at 8–10-fold redundancy, coupled to these sources of contiguity information, enabled not only genome assembly, but also improved quality to about 1 error per 100,000 bases for most of the genome. Additional, focused experiments were performed to fill the gaps or clarify ambiguities.

The Celera effort went straight to paired-end sequencing, eschewing physical maps as an intermediate[31]. An important advance was the transition from greedy algorithms, such as phrap and the TIGR assembler, to the Celera assembler's graph-based approach (overlap–layout–consensus)[22,24,25]. Although Celera had a reasonable strategy for a draft genome, because of the pervasiveness of repetitive sequences, it did not, by itself, result in a high-quality reference, such as the one produced by the HGP's clone-based approach. The current human reference genome descends from the HGP's 2004 product[30], with continuous work by the Genome Reference Consortium to further improve it, including regular releases of reference genome updates[73].

With the advent of NGS in 2005, the number of *de novo* assemblies increased vastly. The seemingly disastrous combination of short reads and repetitive genomes was overcome by new assembly algorithms based on de Bruijn graphs (for example, EULER and Velvet)[74,75]. Nonetheless, particularly when applied to larger genomes and when compared to the genomes of the HGP, their quality was, on average, quite poor. Although shorter read lengths are partly to blame, this is usually overstated. Instead, a principal reason for the poorer quality was the paucity of contiguity

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

BOX 2

The milestones listed below correspond to key developments in the availability of new reference genomes, new sequencing-related computational tools and the applications of DNA sequencing in new ways or to new areas. These are large topics, and we apologize for any omissions.

Genome milestones

1977: Bacteriophage ΦX174 (ref. 72)
1982: Bacteriophage lambda[13]
1995: *Haemophilus influenzae*[26]
1996: *Saccharomyces cerevisiae*[27]
1998: *Caenorhabditis elegans*[28]
2000: *Drosophila melanogaster*[32]
2000: *Arabidopsis thaliana*[146]
2001: *Homo sapiens*[29–31]
2002: *Mus musculus*[147]
2004: *Rattus norvegicus*[148]
2005: *Pan troglodytes*[149]
2005: *Oryza sativa*[150]
2007: *Cyanidioschyzon merolae*[126]
2009: *Zea mays*[151]
2010: Neanderthal[88]
2012: Denisovan[145]
2013: The HeLa cell line[152,153]
2013: *Danio rerio*[154]
2017: *Xenopus laevis*[155]

Computational milestones

1981: Smith–Waterman[156]
1982: GenBank (https://www.ncbi.nlm.nih.gov/genbank/statistics/)
1990: BLAST[16]
1995: TIGR assembler[24]
1996: RepeatMasker
1997: GENSCAN[157]
1998: phred, phrap, consed[22]
2000: Celera assembler[25]
2001: Bioconductor
2001: EULER[74]
2002: BLAT[158]
2002: UCSC Genome Browser[159]
2002: Ensembl[160]
2005: Galaxy[161]
2007: NCBI Short Read Archive
2008: ALLPATHS[162]
2008: Velvet[75]
2009: Bowtie[83]
2009: BWA[82]
2009: SAMtools[84]
2009: BreakDancer[163]
2009: Pindel[164]
2009: TopHat[115]
2010: SOAPdenovo[165]
2010: GATK[85]
2010: Cufflinks[116]
2011: Integrated Genomics Viewer[166]
2013: HGAP/Quiver[167]
2017: Canu[81]

Application milestones

1977: Genome sequencing[72]
1982: Shotgun sequencing[13]
1983, 1991: Expressed sequence tags[107,108]
1995: Serial analysis of gene expression[109]
1998: Large-scale human SNP discovery[168]
2004: Metagenome assembly[122]
2005: Bacterial genome resequencing with NGS[40,41]
2007: Chromatin immunoprecipitation followed by sequencing (ChIP–seq) using NGS[117]
2007–2008: Human genome and cancer genome resequencing using NGS[55,90–92]
2008: RNA-seq using NGS[110–114]
2008: Chromatin accessibility using NGS[118]
2009: Exome resequencing using NGS[97]
2009: Ribosome profiling using NGS[119]
2010: Completion of Phase I of the 1000 Genomes Project[98]
2010: *De novo* assembly of a large genome from short reads[169]
2011: Haplotype-resolved human genome resequencing using NGS[170,171]
2016: Human genome *de novo* assembly with PacBio[172]
2017: Human genome *de novo* assembly with nanopore[64]

methods to complement NGS. Paired-end sequencing was possible with NGS, but *in vitro* library methods were more restricted with respect to the distances that could be spanned. Furthermore, the field lacked 'massively parallel' equivalents of genetic and physical maps.

This 'dark' period notwithstanding, there are good reasons to be optimistic about the future of *de novo* assembly. Firstly, *in vitro* methods that subsample high molecular weight (HMW) genomic fragments, analogous to hierarchical shotgun sequencing, have recently been developed[76,77]. Secondly, methods, such as Hi-C (genome-wide chromosome conformation capture) and optical mapping, provide scalable, cost-effective means of scaffolding genomes into chromosome-scale assemblies[78–80]. Finally, the read lengths of PacBio and ONT sequencing have risen to hundreds of kilobases, and are now more limited by the preparation of HMW DNA than by the sequencing itself. The absence of cloning or amplification steps in single-molecule sequencing pays off, as shown by high-quality PacBio *de novo* assemblies of bacterial genomes with extreme GC content. Long reads have resulted in a re-emergence of strategies used by the Celera assembler, improved to deal with the high error rates and multiple platforms[81]. By combining long reads and even longer-range contiguity information (for example, subsampling HMW DNA, chromatin proximity, optical maps and so on), *de novo* genome assemblies of the quality of the original human reference genome using 'post-Sanger' approaches are finally within sight[73,80].

Genome resequencing

After the HGP, a clear next step was to catalogue genetic variation among humans, that is, 'resequencing'. Because Sanger sequencing costs remained high, resequencing was primarily used to discover common variants, which were then cost-effectively genotyped with microarrays to facilitate genome-wide association studies. The rallying cry for changing this was the 'US$1,000 human genome', the ambitious goal of the resequencing of individuals at a cost nearly one-million-fold below that of assembling the first human genome. The US$1,000 genome concept was discussed as early as 2001 (at the University of California, Santa Cruz Human Genome Symposium (http://genomesymposium.ucsc.edu)), when NGS strategies barely existed, and was formalized a few years later by the Revolutionary DNA Sequencing Technologies program of the National Human Genome Research Institute (NHGRI). The commitment of US$220 million in funding to over 40 academic and 27 commercial entities has helped to drive much of the technological development described above, including direct or indirect support of nearly every successful commercial platform.

Resequencing, that is, mapping sequence reads to a reference genome to identify genetic variants, is a very different task than genome assembly. New algorithms, such as Bowtie and Burrows–Wheeler Aligner (BWA), borrowed concepts from data-compression techniques to enable millions of reads to be efficiently mapped to the reference genome[82,83]. Redundant coverage (for example, 30-fold) is necessary to identify heterozygous

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

variants as well as to distinguish sequencing errors from bona fide variants. Popular packages, initially SAMtools and later GATK, adapted the confidence framework of phred to NGS bases, reads and variants[84,85]. Short reads, particularly when paired, can be uniquely mapped to most of the human genome. But most is not all, and a problem of short-read resequencing is that variants in repetitive regions and structural variants are routinely missed. The extent of this shortcoming is quantified by recent studies that resequence human genomes with PacBio[86]. A second aspect of incompleteness relates to phase relationships between variants in a diploid genome, that is, haplotypes[87]. Fortunately, haplotypes are recovered by many of the same methods that enable contiguity for *de novo* NGS assemblies (and ideally, even *de novo* assemblies would be haplotype-resolved)[77]. Although still not broadly used, these methods are becoming increasingly scalable.

The HGP's human genome was constructed from a mosaic of DNA donors, but mostly derives from one individual, from Buffalo, New York, who had roughly equal parts European and African ancestry[88]. The first individual to have their whole genome resequenced was Craig Venter in 2007, one of the subjects of the Celera genome, which was supplemented with additional data[89]. This was quickly followed in 2008 by the genome of Jim Watson on 454 (ref. 90), and then the genomes of two anonymous individuals[55,91] and the germline and tumour genome of a patient[92] on Solexa/Illumina, and five individuals on Complete Genomics[43]. In this period, whole-genome sequencing (WGS) remained too expensive for most groups to scale, motivating the development of targeted capture methods[93–96] and then whole-exome sequencing (WES), that is, selective sequencing of the 1–2% of the genome that is encodes proteins[97].

As costs approached US$1,000 for WGS[56] and a few hundred dollars for WES, the pace at which individual humans are resequenced has accelerated. The 1000 Genomes Project, launched in 2008, released low-coverage WGS of a few hundred individuals in 2010 and a few thousand individuals in 2015[98,99]. The Exome Sequencing Project released over 6,500 exomes in 2013[100]. The recently released Genome Aggregation Database (http://gnomad.broadinstitute.org/) includes more than 120,000 exomes and over 15,000 genomes. The Genomics England (https://www.genomicsengland.co.uk/), GenomeAsia100K (http://www.genomeasia100k.com/) and NHLBI TOPMed (Trans-Omics for Precision Medicine, https://www.nhlbiwgs.org/) projects each aim to complete WGS on more than 100,000 individuals within the next year or two. Given that these projects represent a fraction of all sequencing being conducted, it is plausible that the genomes of over one million humans have already been resequenced by WES or WGS.

Clinical applications of sequencing

Our ability to sequence human genomes has vastly outpaced our ability to interpret genetic variation, which partly explains why clinical medicine has yet to wholeheartedly embrace WGS. Nonetheless, there are some areas in which DNA sequencing is already proving clinically useful, three of which we highlight here.

The most unexpected area of the clinical impact of DNA sequencing has been non-invasive prenatal testing (NIPT, see Fig. 2). Pioneering studies by Lo and Quake in 2008 have demonstrated that the simple counting of DNA fragments released into the maternal circulation by a fetus during pregnancy can detect chromosomal aneuploidies[101,102]. Screening tests that were based on this strategy were adopted faster than any molecular test in history, and several million pregnant women around the world have already benefited from low-pass WGS for NIPT.

An early application of WES was to rapidly discover new genes for, and to diagnose patients affected by, Mendelian disorders[97,103]. This was quickly followed by the discovery that substantial proportions of neurodevelopmental disorders are attributable to *de novo* mutations in coding sequences[104]. WES is increasingly used as a primary tool for diagnosing Mendelian and neurodevelopmental disorders, particularly in paediatric populations, with the rate of diagnosis of patients with suspected Mendelian disease by WES at 25% and rising[105].

Our understanding of cancer, fundamentally a disease of the genome, is gradually being transformed by DNA sequencing. Large-scale resequencing has laid bare the extraordinary genetic heterogeneity of cancers, effectively defining a molecular taxonomy[106]. DNA sequencing is impacting clinical cancer care by: (1) suggesting targeted therapies, based on the mutations present in an individual cancer; (2) enabling non-invasive diagnosis or monitoring by sequencing of tumour-released circulating cells or cell-free DNA; (3) identifying cancer-specific, protein-altering mutations that may serve as neoantigens for 'personal vaccines'. Although, the success stories in each of these areas are still few and far between, relative to the overall burden of cancer, progress is clearly being made.

Sequencers as a molecule counting device

While 'expressed sequenced tags'[107] were considered a shortcut to gene discovery as early as 1983[108], it was SAGE (serial analysis of gene expression; 1995) that introduced the idea of sequencing to 'digitally quantify' gene expression[109]. SAGE concatenated cDNA-derived tags for Sanger sequencing, with tags that are just long enough to map to a gene. As early as 2000, Brenner and Lynx Therapeutics demonstrated 'massively parallel signature sequencing' of cDNA tags, an important forerunner of NGS[47]. However, this concept was not widely adopted until the development of RNA sequencing (RNA-seq) by five groups in 2008. RNA-seq uses NGS to quantify and characterize the transcriptome by shotgun sequencing of either full-length or 3′ ends of cDNA[110–114]. RNA-seq has marked advantages over microarrays, the most notable of which is that transcript counts lead to straightforward statistics relative to analogue, hybridization-based signals, facilitated by new software packages, such as TopHat and Cufflinks[115,116].

Also around 2008, small laboratories that were early adopters of NGS developed 'digital quantification' methods for transcription-factor binding[117], chromatin accessibility[118] and translation[119]. In the following decade, hundreds of protocols were developed that facilitate the use of DNA sequencing as a 'molecule counter' for the characterization of a remarkable range of biochemical or molecular phenomena, including transcription, translation, DNA replication, the secondary structure of RNA, chromosome conformation, nucleic-acid modifications, post-translational modifications, nucleic acid–protein interactions and protein–protein interactions. These are catalogued in other reviews and resources (ref. 120 and http://enseqlopedia.com/).

The use of sequencers as molecule-counting devices was immediately immensely popular, and probably had a larger role than assembly or resequencing in driving the widespread adoption of NGS in biomedical research. DNA sequencers are increasingly to the molecular biologist what a microscope is to the cellular biologist—a basic and essential tool for making measurements. In the long run, this may prove to be greatest impact of DNA sequencing.

Metagenome sequencing

Shotgun sequencing of complex communities of microorganisms[121–123], for example, metagenome sequencing of environmental or human microbiomes, has emerged as a field of its own, bringing with it unique challenges with respect to assembly, resequencing and counting. Other reviews have recently covered this topic[124,125].

The future of DNA sequencing

In the long view of scientific history, DNA sequencing remains a young technology. Here, we briefly consider its future in a few existing or emerging areas.

Genome diversity

A 100% complete genome, that is, the telomere-to-telomere sequence for each chromosome with no gaps or ambiguities, has been achieved for possibly only one eukaryote so far[126]. As sequencing technologies continue to evolve, we are optimistic that we will resolve challenging regions of additional genomes (for example, centromeres). There are

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

millions of living species on earth (and far more extinct species), each with a genome waiting to be sequenced, as well as countless microbiomes and metagenomes. A comprehensive view of genomic diversity may prove useful in surprising ways, for example, for protein structure determination[127].

Population-scale resequencing

We are approaching the milestone where approximately 0.1% of living humans will have had their genomes resequenced to some degree, while resequencing of the genomes of our ancestors and other hominins is reshaping our understanding of human history[88]. The number of *de novo* point mutations occurring in recent generations vastly exceeds the number of nucleotides in the human genome. Eventually, aggregating tens of millions of genomes may enable a nucleotide-level footprint of the human genome (that is, observing all heterozygous variants compatible with life). DNA sequencing also is increasingly useful for forensics, without necessarily requiring a sample from the identified individual[128].

Developmental biology

We each develop from a single cell into a highly organized mass of trillions of cells. However, our understanding of development remains coarse. Recent technologies enable scalable, sequencing-based profiling of single cells. Although popular approaches are *ex vivo* (for example, single-cell RNA-seq), a more radical approach is to perform RNA or protein sequencing *in situ*, thereby retaining the spatial context[129,130]. Other emerging strategies use *in vivo* genome editing to track cell-lineage relationships[131] or transport barcodes to catalogue neuronal connections[132]. Editing of DNA can potentially be used to record biological events more generally, for example, to monitor gene expression[133] or calcium[134].

Real-time, portable sensors

Nanopore sequencers currently have a mass of 70 g and yield data within 30 min of sample application. One can imagine disseminated networks of nanopore sequencers enabling 'universal monitoring' of nucleic acids, in environmental settings and in everyday human life, for example, fine-grained tracking of our air, food and body, potentially streaming data from millions of devices and integrating with GPS and audio-visual data.

Unconventional uses

DNA-sequencing technologies will probably prove useful in additional, surprising ways. For example, NGS has recently been used to recover large amounts of data encoded in synthetic DNA[135]. Nanopores may find uses beyond sequencing, for example, for monitoring analyte binding[136], chemical nanomachines[137] or protein folding/unfolding[138].

DNA sequencing as the new microscope

It has been about 400 years since the invention of light microscopy, a technology which continues to be used and to evolve. By comparison, it has been only 40 years since the invention of DNA sequencing; the technologies for which are likely to also continue to develop in the coming decades and centuries. On the basis of how quickly it has transformed biomedical research, and is beginning to transform clinical medicine, we predict that DNA sequencing will have a longevity and impact on par with or exceeding that of the microscope.

Received 13 July; accepted 21 September 2017.
Published online 11 October 2017.

1. Sanger, F. Sequences, sequences, and sequences. *Annu. Rev. Biochem.* **57,** 1–28 (1988).
2. Sanger, F. Nobel lecture: the chemistry of insulin. https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1958/sanger-lecture.html (2017).
3. Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4,** 283–293 (1950).
4. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* **147,** 1462–1465 (1965).
5. Sanger, F., Brownlee, G. G. & Barrell, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* **13,** 373–398 (1965).
6. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35,** 523–537 (1968).
7. Gilbert, W. & Maxam, A. The nucleotide sequence of the lac operator. *Proc. Natl Acad. Sci. USA* **70,** 3581–3584 (1973).
8. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74,** 5463–5467 (1977).
Refs 8, 9: The seminal papers by Sanger, Nicklen & Coulson and Maxam & Gilbert describing the first widely adopted methods for DNA sequencing.
9. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
10. Maniatis, T., Jeffrey, A. & van deSande, H. Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis. *Biochemistry* **14,** 3787–3794 (1975).
11. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6,** 2601–2610 (1979).
12. Messing, J., Crea, R. & Seeburg, P. H. A system for shotgun DNA sequencing. *Nucleic Acids Res.* **9,** 309–321 (1981).
13. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA. *J. Mol. Biol.* **162,** 729–773 (1982).
14. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).
15. Connell, C. *et al.* Automated DNA sequence analysis. *Biotechniques* **5,** 342–348 (1987).
16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
17. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341 (1987).
18. Tabor, S. & Richardson, C. C. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl Acad. Sci. USA* **84,** 4767–4771 (1987).
19. Craxton, M. Linear amplification sequencing, a powerful method for sequencing DNA. *Methods* **3,** 20–26 (1991).
20. DeAngelis, M. M., Wang, D. G. & Hawkins, T. L. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23,** 4742–4743 (1995).
21. Zhang, J. *et al.* Use of non-cross-linked polyacrylamide for four-color DNA sequencing by capillary electrophoresis separation of fragments up to 640 bases in length in two hours. *Anal. Chem.* **67,** 4589–4593 (1995).
22. Green, P. phred, phrap, consed. http://www.phrap.org/phredphrapconsed.html (2017).
phred introduced quantitative, reliable metrics for base quality, substituting human judgement with computers, a process that occurred repeatedly over the course of the HGP.
23. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6,** 593–608 (1990).
24. Sutton, G. G., White, O., Adams, M. D. & Kerlavage, A. R. TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.* **1,** 9–19 (1995).
25. Myers, E. W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287,** 2196–2204 (2000).
The Celera assembler introduced an overlap–layout–consensus approach to deal with the problems posed by repeats and the millions of reads needed to produce a reliable assembly.
26. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
27. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
28. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
29. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
Refs 29–31: The HGP and Celera produced draft sequences of the human genome with the HGP later publishing a more complete, relatively error-free reference.
30. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
31. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
32. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
33. Balasubramanian, S., Klenerman, D. & Barnes, C. Arrayed polynucleotides and their use in genome analysis. Patent US20030022207 (2003).
34. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S. R. Sequence information can be obtained from single DNA molecules. *Proc. Natl Acad. Sci. USA* **100,** 3960–3964 (2003).
35. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320,** 106–109 (2008).
36. Adams, C. P. & Kron, S. J. Method for performing amplification of nucleic acid with two primers bound to a single solid support. Patent US5641658 (1997).
37. Chetverina, H. V. & Chetverin, A. B. Cloning of RNA molecules *in vitro*. *Nucleic Acids Res.* **21,** 2349–2353 (1993).
38. Mitra, R. D. & Church, G. M. *In situ* localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res.* **27,** e34–e39 (1999).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

致谢

四年的本科学习即将结束，在临近毕业之时，忽而对学习这件事本身产生了全新的认识。我相信这一份学历给我的，是又一个新的开始。在此，特向在学习、生活过程中帮助过我的老师、同学、朋友、亲人们表示衷心的感谢。

首先要感谢本科学习过程中的各位老师，尤其是王丽芝老师。感谢老师将我的视野提升到了一个新的高度，王老师在学习中重点分明、思路清晰的品质让我受益匪浅。感谢老师对我实验方面的悉心指导，更感谢老师将我推荐到药用植物研究所来做毕业设计，在这里我感受到了和大学中完全不同的氛围，让我真正体会到了什么是科研。感谢药用植物研究所的孙超老师、李滢老师在毕业设计中的指导，以及对我实习生活中的关照。

感谢中药资源教研组的各位老师，我从马琳老师身上感受到对于植物真挚的喜爱，这种感情在课堂上深深地感染了我；感谢李先宽老师在长白山上的言传身教；感谢王海英老师在学习和大创课题的耐心指导；感谢张坚老师和向蓓蓓老师对于专业课的教导。

感谢中药资源与开发专业 2014 级全体同学，感谢 214 的另外三名室友，感谢匡雪君、刘琬菁、褚丽华、梁彤彤、张建红等师姐，感谢任凤鸣、徐志超、浦香东等师兄，感谢 2012、2013 级给予过帮助的学长学姐们。

感谢天津中医药大学与北京药用植物研究所提供的学习与科研环境。感谢父母对于生活资金的支持。

前期工作基础：测序数据由北京诺禾致源公司完成，二代数据混合拼接由李滢老师完成。