申请代码	F020504
受理部门	
收件日期	
受理编号	6177051641

国家自然科学基金

申 请 书

（2 0 1 7 版）

资助类别： 面上项目

亚类说明：

附注说明： 常规面上项目

项目名称： 多倍体单体型从头组装算法研究

申 请 人： 谢民主 **电 话：** 13187021287

依托单位： 湖南师范大学

通讯地址： 湖南长沙湖南师范大学物理与信息科学学院

邮政编码： 410081 **单位电话：** 0731-88872262

电子邮箱： xieminzhu@hotmail.com

申报日期： 2017年02月08日

国家自然科学基金委员会

基本信息

申请人信息	姓　　名	谢民主	性别	男	出生年月	1969年10月	民族	汉族
	学　　位	博士	职称	教授		每年工作时间（月）		10
	电　　话	13187021287		电子邮箱		xieminzhu@hotmail.com		
	传　　真			国别或地区		中国		
	个人通讯地址	湖南长沙湖南师范大学物理与信息科学学院						
	工　作　单　位	湖南师范大学/物理与信息科学学院						
	主要研究领域	生物信息学，算法设计与分析						
依托单位信息	名　　称	湖南师范大学						
	联　系　人	王青春		电子邮箱		kjch@hunnu.edu.cn		
	电　　话	0731-88872262		网站地址		http://kjc.hunnu.edu.cn/EntranceIndex.action		
合作研究单位信息	单位名称							
	清华大学							
项目基本信息	项目名称	多倍体单体型从头组装算法研究						
	英文名称	Research on algorithms of de novo single individual haplotyping for polyploids						
	资助类别	面上项目		亚类说明				
	附注说明	常规面上项目						
	申请代码	F020504.生物信息计算		C060702.生物信息算法及工具				
	基地类别							
	研究期限	2018年01月01日 -- 2021年12月31日		研究方向：生物信息学/计算生物信息学				
	申请直接费用	71.0000万元						
中文关键词		生物信息学；单体型；DNA序列分析；新一代测序技术；算法						
英文关键词		Bioinformatics; haplotype; DNA sequences analysis; next-generation sequencing technology; algorithm						

中文摘要	一些重要经济作物的体细胞有多套染色体组，利用测序技术确定每套染色体的DNA序列，即单体型组装，对作物改良有重要意义。当前多倍体单体型组装算法需要把测序读段联配到参考基因组上，无法应用于参考基因组未知的多倍体。 本课题研究无需参考基因组序列、融合多平台测序数据的多倍体单体型从头组装算法。本课题将基于De Bruijn图在错误率低的短读段数据上构建可信度高的contig集，通过基于contig的联配，用聚类分析、机器学习等手段进行读段纠错；用重叠-扩展方式对contig及相关读段进行局部组装，在长读段、paired-end读段、局部组装和contig数据基础上针对多倍体植物基因组重复率高、各单体型对之间差异率明显不同的特征建立优化模型；结合参数计算、动态规划和高级图算法等技术设计快速有效求解算法，构建各单体型的DNA序列。高效精确的多倍体单体型从头组装算法将促进多倍体生物全基因组测序工程。
英文摘要	Some economically important plants have more than two copies of each chromosome. Constructing the DNA sequence of each copy of chromosome by DNA sequencing, i.e. haplotype assembly, is valuable to crop improvement. Current algorithms of haplotype assembly for polyploids have to align the DNA reads to reference genomes and cannot be applied to the polyploids without known reference genomes. The project focuses on algorithms of haplotype assembly for polyploids without known reference genomes based on DNA reads from multiple sequencing platforms. The project will use a De Bruijn graph to construct high reliable contigs from short reads with a low error rate, and use contig-based alignment, cluster analysis and machine learning to correct the sequencing errors of reads. Then local assemblies of the contigs and the related reads will be conducted via overlap-extension approaches. Based on the information provided by long reads, paired-end reads, local assemblies and contigs, we will propose haplotype assembly optimal models to deal with the challenge caused by the large and rich repeats of polyploid plant genomes and the fact that difference ratios varies much between different haplotype pairs. Furthermore, we will combine parameterized computation, dynamic programming, advanced graph algorithm and other algorithm designing techniques, design fast and effective algorithms to solve the optimal models, and construct the DNA sequence for each haplotype. Efficient and accurate algorithms of de novo haplotype assembly will prompt genome sequencing of polyploids.

项目组主要参与者（注：项目组主要参与者不包括项目申请人）

编号	姓名	出生年月	性别	职 称	学 位	单位名称	电话	电子邮箱	证件号码	每年工作时间（月）
1	姜涛	1963-05-06	男	教授	博士	清华大学	13910065021	jiang@cs.ucr.edu	HB675861	1
2	钟坚成	1981-12-06	男	副教授	博士	湖南师范大学	13207316286	superzjc@163.com	430105198112060512	4
3	周建宇	1993-02-15	男	博士生	学士	清华大学	15201376280	sherry521007@gmail.com	130903199302150337	10
4	熊袁鹏	1994-07-28	男	博士生	学士	清华大学	13397040728	yuanpengxiong@foxmail.com	360425199407282516	10
5	叶云洋	1982-10-14	男	博士生	硕士	湖南师范大学	15073224060	85372274@qq.com	36232519821014295X	10
6	彭哲也	1992-08-24	男	硕士生	学士	湖南师范大学	15576633654	394053765@qq.com	430922199208240010	10
7	周佩霞	1990-02-09	女	硕士生	学士	湖南师范大学	13755112957	344919395@qq.com	430124199002094627	10
8	唐紫珺	1993-05-31	女	硕士生	学士	湖南师范大学	18974663735	772452121@qq.com	431103199305310023	10
9	喻昕	1993-12-08	女	硕士生	学士	湖南师范大学	18570341514	274847928@qq.com	430124199312084985	10

总人数	高级	中级	初级	博士后	博士生	硕士生
10	3				3	4

国家自然科学基金项目资金预算表

项目申请号：6177051641 项目负责人：谢民主 金额单位：万元

序号	科目名称	金额	备注
	(1)	(2)	(3)
1	**一 、直接费用**	71.0000	
2	1、 设备费	10.9000	
3	(1)设备购置费	8.50	购买必要的计算和数据存储设备
4	(2)设备试制费	0.0000	
5	(3)设备改造与租赁费	2.40	现有计算设备升级改造
6	2、 材料费	3.80	项目实施过程中涉及一些原材料和耗材
7	3、 测试化验加工费	1.60	计算资源使用费用等
8	4、 燃料动力费	0.00	
9	5、 差旅/会议/国际合作与交流费	7.10	用于团队成员参加国内外学术交流等
10	6、 出版/文献/信息传播/知识产权事务费	12.80	审稿费、版面费、网络通信费等
11	7、 劳务费	28.00	研究生的劳务费
12	8、 专家咨询费	4.00	领域专家咨询费
13	9、 其他支出	2.80	项目组研讨会所需的其他耗材费用
14	**二、 自筹资金来源**	0.0000	

预算说明书（定额补助）

（请按《国家自然科学基金项目资金预算表编制说明》中的要求，对各项支出的主要用途和测算理由及合作研究外拨资金、单价≥10万元的设备费等内容进行详细说明，可根据需要另加附页。）

1、设备费：

（1）设备购置费

每年计划招收博士研究生1名、硕士研究生2名参与项目，四年预计共需新购6台计算机，费用 0.6万/台*6台 = 3.6万；购置一台工作站用于运行和测试对计算资源要求较高的算法，预计经费1.6万；购置一台数据服务器下载和存储大量DNA测序数据，预计经费1.8万；为研究生购置打印机一台，预算1万；购置扫描仪一台，预算0.5万；小计：3.6+1.6+1.8+1+0.5 = 8.5万。

（3）设备改造与租赁费

已有的计算机内存、硬盘等部件升级，维修，预计经费0.6万/年*4年 = 2.4万。

2、材料费

项目实施过程中涉及一些原材料和耗材，如数据存储器材、计算机、网络通信与打印机耗材等。移动硬盘10个，预算：0.08 *10 = 0.8万；键盘、鼠标、网络通信设备等材料更换，四年预算1.6万；硒鼓四年预算0.8万；墨盒四年预算0.6万；

小计：0.8+1.6+0.8+0.6= 3.8万（其中外拨合作单位清华大学0.8万）。

3、测试化验加工费：主要用于计算中心算法测试计算资源使用经费等，四年预算1.6万。

5、差旅/会议/国际合作与交流费

（1）四年内计划参加国内交流研讨会议10人次，0.38万/人次*10人次= 3.8万；

（2）四年内计划参加国际会议3次，每次预算1.1万(其中预计每人次会议注册费用0.4万，往返机票、住宿费用0.7万)，共需3.3万；

小计：7.1万（其中外拨合作单位清华大学2.5万）。

6、出版/文献/信息传播/知识产权事务费

（1）国际、国内期刊发表论文12-16篇，审稿费、版面费预算10万（其中外拨合作单位清华大学3万）；

（2）专利、软件著作权申请、查新、文献检索，四年预算2万（其中外拨合作单位清华大学0.4万）；

（3）网络通信费四年预算0.8万（其中外拨合作单位清华大学0.4万）；

小计：12.8万（其中外拨合作单位清华大学3.8万）。

7、劳务费

课题执行4年中预计平均每年参与课题的博士研究生4名，硕士研究生6名

（1）博士研究生劳务费预算：1万/人年*4人*4年 = 16万（其中外拨合作单位清华大学博士研究生劳务费8万）；

（2）硕士研究生劳务费预算：0.5万/人年*6人*4年 = 12万；

小计：28万（其中外拨合作单位清华大学8万）。

8、专家咨询费

预计咨询生物、数学领域专家5人次，每次咨询费用预算0.8万，总计4万。

9、其他支出

主要用于购买酒精笔、项目组实验室日常运行、科研研讨活动相关设施维护费用，预计每年0.7万，四年总计2.8万（其中外拨合作单位清华大学0.4万）。

总计：71万（其中外拨合作单位清华大学15.5万）。

报告正文

参照以下提纲撰写，要求内容翔实、清晰，层次分明，标题突出。**请勿删除或改动下述提纲标题及括号中的文字。**

（一）立项依据与研究内容（4000-8000 字）：

1. 项目的立项依据（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）：

1.1 研究意义

农作物是当前人类食物的主要来源，其育种及品种改良对世界粮食安全有重要意义。一些广为种植的农作物如小麦、棉花、甘蔗等均为多倍体[1]，即其体细胞拥有两套以上的染色体组（基因组）；**多倍体特征**（polyploidy）长期以来被认为是导致植物表型多样化的重要原因，了解每套染色体的 DNA 序列（haplotype，**单体型**）将有力地促进植物进化史构建和作物的改良[2, 3]。

随着 DNA 测序技术的快速发展，新的测序仪测得的 DNA 读段长度不断增长，而测序成本快速下降[4, 5]，对多倍体植物进行全基因组测序，从 DNA 读段重构每套染色体组的 DNA 序列日益可行[6, 7]。可是 DNA 测序的固有误差，植物基因组重复率高和多倍体特性使多倍体植物的全基因组测序和单体型重建面临巨大挑战。

自从 2000 年拟南芥被测序以来，水稻、玉米、大豆等经济植物的全基因组参考序列陆续发布[8]，可是复杂多倍体植物的全基因组测序却进展缓慢。异源四倍体陆地棉（*G. hirsutum L.*）是世界上最重要的纤维作物和模式多倍体作物，中国农业科学院棉花研究所联合北京大学等多家科研单位历经 7 年多的时间，于 2015 年才发表其基因组序列草图[9]。国际小麦基因组测序联盟（The International Wheat Genome Sequencing Consortium, IWGSC）从 2005 年就开始致力于普通小麦的全基因组测序，2014 年 7 月才发布异源六倍体普通小麦（*Triticum aestivum*）“中国春”的基于染色体的基因组序列草图[10]，2017 年 1 月其第一版高质量参考基因组序列[11]得以发表。该项工作采用生物

实验技术染色体分选（chromosome sorting）把各染色体分离，然后分别测序组装，由几十家世界知名的科研院所花费巨资协作完成。

因此**研究如何利用有效的计算技术融合新一代不同 DNA 测序平台产生的大量数据，花费较少的人力和物力来确定多倍体的单体型**具有极其重要的现实意义。

1.2 国内外研究现状及发展动态分析

高等生物的体细胞通常含有不止一套染色体，测定其中一套染色体的 DNA 序列，实验室最直接的方法是利用如染色体分选等生物实验技术把染色体分离，然后进行测序[12]，国际小麦基因组测序联盟对普通小麦的测序就是采用这样的方法[10]，可是这些方法技术难度大，所需时间长、费用高[13]。中国农业科学院棉花研究所整合了全基因组鸟枪测序、BAC-to-BAC、高密度遗传图谱构建等策略，对陆地棉进行了全基因组测序和组装[9]，BAC-to-BAC 和高密度遗传图谱构建同样需要大量复杂的生物实验室工作。Chapman 等采用全基因组鸟枪测序策略对异源六倍体小麦“Synthetic W7984”和“Opata M85”进行测序，可其组装必须依靠一个超稠密遗传图谱（ultra-dense genetic map），这个图谱是通过对它们 90 个二倍体后代进行测序后构建的[14]，需要**大量的育种和额外的测序工作**。综上所述，**这些方法需要大量繁重且难度大的生物实验操作或大量额外的测序，难以广泛推广。**

下面仅对利用计算技术，在测序仪产生的、来自细胞多套染色体的 DNA 读段混合数据集的基础上，进行单体型重建的研究进行简述。

1.2.1 二倍体单体型组装

利用计算技术在 DNA 测序读段数据的基础上构建人类等二倍体个体的单体型目前得到了广泛研究[15-18]。由于人类参考基因组序列（来自多个不同个体的基因组表决序列）已经获得，人类个体的单体型组装算法通常是把 DNA 读段与参考基因组序列进行联配，提取出读段的多态性位点信息（通常是**单核苷酸多态性 SNP** 信息），在此基础上进行单体型重建。

2001 年，Lancia 等最先对该问题进行研究，提出了最少片段删除(Minimum Fragment Removal，MFR)、最少 SNP 位点删除(Minimum SNP Removal，MSR)和最长单体型重建(Longest Haplotype Reconstruction，LHR)三个

计算模型[19]。2002 年，Lippert 等提出了最少错误更正模型(Minimum Error Correction，MEC) [20]；2010 年 Duitama 等提出了最大片段割模型(Maximum Fragments Cut，MFC)[21]；2012 年项目申请者**谢民主**等提出了平衡优化分区模型(Balanced Optimal Partition，BOP)[22]。

上述模型及其扩展均被证明是 NP-难的[21-26]，求解最优解的精确算法的复杂度至少与输入数据的某个参数特征成指数增长[25-30]：例如谢民主等在 ISMB 2008 上提出求解 WMLF/GS 模型(Weighted Minimum Letter Flips with GenoSpectrum，为 MEC 模型的扩展)的精确算法，其时间复杂度随读段覆盖的多态性位点个数的最大值成指数增长[26]；2016 年，Pirola 等在 Bioinformatics 上发表论文提出了一个精确算法求解 MEC 模型，其算法复杂度随一个多态性位点上需要纠正的测序错误数的最大值成指数增长[27]。

上述精确算法对于在特定参数上取值小的输入数据能快速获取模型的最优解，但不适于复杂的输入数据，因此求解这些组合优化模型的大量启发式算法被提出[21, 22, 31-34]。

为求解 MEC 模型，Zhao 等提出了动态聚类算法[33]，王瑞省等提出了遗传算法[34]， Qian 等提出了粒子群优化算法[35]，吴璟莉等提出了单亲遗传算法[36]，谢民主等提出了一个基于两位点连锁图的启发式算法[37]，Chen 等则提出基于整数线性规划结合启发式规则对 MEC 求解[38]；而 Bansal 等则把 MEC 转化为最大割问题，采用两步贪婪策略，设计了算法 HapCut 求解 MEC 模型[39]，该算法曾用于一个印度人的单体型组装[40]。

Duitama 等采用与 HapCut 类似的贪婪策略，设计了算法 ReFHap 求解 MFC 模型[21]，该算法曾用于个体 NA12878(来自国际人类基因组单体型图计划招募的 CEU 群体) [41]、德国人 MP1[42]和中国人 YH[43]的单体型组装；J. Craig Venter 的单体型也是采用贪婪策略在一个初始解的基础上不断迭代进行组装[44]。谢民主等把 top-*k* 的策略和动态规划结合起来，设计了求解 BOP 模型的启发式算法[22]，其单体型重建精度比 ReFHap 有所改善，速度显著提高。

除了上述组合计算模型以外，Li 等[45]和 Kuleshov 等[46, 47]提出了概率模型，在给定个体 DNA 片段数据的基础上，求出具有最大条件

概率的一对单体型。由于单体型对的可能数目随着单体型的多态性位点成指数增长，因此目前求解该概率模型的算法主要是基于期望最大化(Expectation-Maximization，EM)、 Markov 链、Gibbs 取样等技术的启发式算法[45-50]。2016 年，Cai 等则把二倍体单体型组装问题转化成矩阵分解问题，采用梯度下降法进行求解[51]。

1.2.2 多倍体单体型组装

k 倍体($k > 2$)单体型组装比二倍体单体型组装具有更大的挑战性，对于一个有 *n* 个异构多态性位点的 *k* 倍体，其基因型（genotype, *k* 倍体在多态性位点上 *k* 个取值集合的序列）取值空间为 $(k\text{-}1)^n$，其 *k* 个单体型的组合空间至少 $2^n(k\text{-}1)^n$ [52, 53]。对二倍体单体型组装算法的简单扩展通常无法应用于多倍体组装。

最近有学者开始对该问题进行研究，提出了几个多倍体单体型组装算法。2013 年，Aguiar 等提出了算法 HapCompass[54]，该算法用一个图表示联配后的 DNA 读段数据，图的顶点表示 SNP，边表示有读段支持边两端的 SNP 取值出现在同一个单体型上；然后把单体型组装问题转化为图的生成树问题；HapCompass 通过基于环的局部优化方法试图寻找一个去掉的边的权重和最小的生成树。

2014 年，Berger 等提出了算法 HapTree[53]，HapTree 基于最大似然度估计（maximum-likelihood estimation）对 *k* 倍体单体型进行重建，试图发现能以最大可能性生成输入 DNA 读段数据的 *k* 个单体型。为了降低计算复杂度，HapTree 采用了与动态规划相似的策略，首先发现在前面少数几个 SNP 位点上具有较高似然度的 *k*-单体型集合，然后接着往下一个 SNP 位点扩展。

2015 年，Das 等提出了算法 SDhaP[55]，SDhaP 把 *k* 倍体单体型组装转化为一个半定规划（semidefinite programming, SDP）问题，然后采用拉格朗日松弛（Lagrangian relaxation）、随机投影和贪婪策略进行求解。

2016 年，谢民主等把多倍体单体型组装问题建模成 DNA 读段的多分区优化模型，通过限制动态规划迭代过程中中间解的个数，设计了能平衡求解精度和求解速度的两个算法 H-PoP 和 H-PoPG[52]。Puljiz 等则把单体型组装问题类比成从有噪通道接收信号（DNA 读段数据）

的解码问题，采用通信系统中的置信度传播算法（belief propagation）求解多倍体单体型组装问题[56]。

上述的单体型组装算法的输入数据均需把 DNA 读段联配到参考基因组序列上，提取出读段在多态性位点上的取值，无法用于参考基因组序列未知的物种。对于**同源多倍体**，其染色体组均是来自同一物种，各单体型 DNA 序列相差很小，当前有大量 DNA 序列组装软件[57-59]如 Celera 组装器[60]、Velvet[61]、EULER-SR[62]、ABySS [63]、SOAPdenovo[64]和 hybridSPAdes[65]等可用于从 DNA 读段数据组装出多倍体基因组表决序列，然后在此基础上，把 DNA 读段联配到表决序列后，再进行多倍体单体型组装。然而很多多倍体植物属于**异源多倍体**，如普通小麦、油菜、棉化、烟草等，是不同物种杂交产生的后代经过染色体加倍而成的，来自不同物种的染色体 DNA 序列差异甚大，**以上组装软件无法对异源多倍体的全基因组测序读段数据进行有效组装。**而**无需参考基因组序列，直接基于全基因组测序读段数据的多倍体单体型从头组装算法还未看到有相关文献报道。**

因此，研究如何**综合利用不同新一代测序平台生成的 DNA 读段数据，设计有效的计算模型和高效的计算机算法，在参考基因组序列未知的条件下，以较少的人力和物力确定多倍体各单体型**，对于推动复杂多倍体植物全基因测序有着重要的意义。

主要参考文献目录

[1] A.R. Leitch, I.J. Leitch, Genomic plasticity and the diversity of polyploid plants. *Science*, **2008**, 320(5875): 481-3.

[2] S. Renny-Byfield, J.F. Wendel, Doubling down on genomes: Polyploidy and crop plants. *Am J Bot*, **2014**.

[3] 张学勇, 马琳, 郑军, 作物驯化和品种改良所选择的关键基因及其特点. **作物学报**, **2017**, 43(2): 157-170.

[4] Y. Feng, et al., Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics*, **2015**, 13(1): 4-16.

[5] M. Bahassi el, P.J. Stambrook, Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, **2014**, 29(5): 303-10.

[6] M. Thudi, et al., Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics*, **2012**, 11(1): 3-11.

[7] R. Ming, C.M. Wai, Assembling allopolyploid genomes: no longer formidable. *Genome Biology*, **2015**, 16.

[8] M.E. Bolger, et al., Plant genome sequencing - applications for crop improvement. *Curr Opin Biotechnol*, **2014**, 26: 31-7.

[9] F. Li, et al., Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. *Nat Biotechnol*, **2015**, 33(5): 524-30.
[10] The International Wheat Genome Sequencing Consortium, A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **2014**, 345(6194): 1251788.
[11] The International Wheat Genome Sequencing Consortium, IWGSC Reference Sequence v1.0 assembly now available at URGI. 2017 [cited 2017 Jan.18]; Available from: http://www.wheatgenome.org/News/Latest-news/RefSeq-v1.0-URGI.
[12] H. Yang, X. Chen, W.H. Wong, Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci U S A*, **2011**, 108(1): 12-7.
[13] D. Porubsky, et al., Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res*, **2016**, 26(11): 1565-1574.
[14] J.A. Chapman, et al., A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol*, **2015**, 16: 26.
[15] M. Xie, et al., Computational Models and Algorithms for the Single Individual Haplotyping Problem. *Current Bioinformatics*, **2010**, 5(1): 18-28.
[16] X. Zhang, et al., Models and algorithms for haplotyping problem. *Current Bioinformatics*, **2006**, 1: 105 - 114.
[17] S.R. Browning, B.L. Browning, Haplotype phasing: existing methods and new developments. *Nat Rev Genet*, **2011**, 12(10): 703-14.
[18] J.-K. Rhee, et al., Survey of computational haplotype determination methods for single individual. *Genes & Genomics*, **2016**, 38(1): 1-12.
[19] G. Lancia, et al., SNPs problems, complexity and algorithms. *Proc. Ann. European Symp. on Algorithms (ESA), Volume 2161 of Lecture Notes in Computer Science*, **2001**: 182 - 193.
[20] R. Lippert, et al., Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem. *Brief Bioinform*, **2002**, 3(1): 23-31.
[21] J. Duitama, et al., ReFHap: a reliable and fast algorithm for single individual haplotyping, in Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, 2010, ACM: Niagara Falls, New York, p. 160-169.
[22] M. Xie, J. Wang, T. Jiang, A fast and accurate algorithm for single individual haplotyping. *BMC Systems Biology*, **2012**, 6(Suppl 2): S8.
[23] R. Cilibrasi, et al., The complexity of the single individual SNP haplotyping problem. *Algorithmica*, **2007**, 49(1): 13 - 36.
[24] Y. Xie, et al., Meranzin hydrate exhibits anti-depressive and prokinetic-like effects through regulation of the shared alpha 2-adrenoceptor in the brain-gut axis of rats in the forced swimming test. *Neuropharmacology*, **2012**, 67C: 318-325.
[25] V. Bafna, et al., Polynomial and APX-hard cases of the individual haplotyping problem. *Theoretical Computer Science*, **2005**, 335: 109 - 125.
[26] M. Xie, J. Wang, J. Chen, A model of higher accuracy for the individual haplotyping problem based on weighted SNP fragments and genotype with errors. *Bioinformatics*, **2008**, 24(13): i105-13.
[27] Y. Pirola, et al., HapCol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, **2016**, 32(11): 1610-7.
[28] P. Bonizzoni, et al., On the Fixed Parameter Tractability and Approximability of the Minimum Error Correction Problem, in 26th Annual Symposium on Combinatorial Pattern Matching (CPM), C. Ferdinando and e. al., Editors, 2015, Springer International Publishing: Ischia, Italy, p. 100–113.

[29] D. He, et al., Optimal algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, **2010**, 26(12): i183-90.

[30] 谢民主, 陈建二, 王建新, 个体单体型问题参数化算法研究. **计算机学报**, **2009**, 32(8): 1637-1650.

[31] L.M. Genovese, F. Geraci, M. Pellegrini, SpeedHap: an accurate heuristic for the single individual SNP haplotyping problem with many gaps, high reading error rate and low coverage. *IEEE/ACM Trans Comput Biol Bioinform*, **2008**, 5(4): 492-502.

[32] A. Panconesi, M. Sozio, Fast hare: a fast heuristic for single individual SNP haplotype reconstruction. *Proc. WABI, Volume 3240 of Lecture Notes in Computer Science*, **2004**: 266 - 277.

[33] Y.Y. Zhao, et al., Haplotype assembly from aligned weighted SNP fragments. *Comput Biol Chem*, **2005**, 29(4): 281-7.

[34] R.S. Wang, et al., Haplotype reconstruction from SNP fragments by minimum error correction. *Bioinformatics*, **2005**, 21(10): 2456-62.

[35] W. Qian, et al., Particle swarm optimization for SNP haplotype reconstruction problem. *Applied Mathematics & Computation*, **2008**, 196(1): 266-272.

[36] J. Wu, J. Wang, J.E. Chen, A parthenogenetic algorithm for single individual SNP haplotyping. *Engineering Applications of Artificial Intelligence*, **2009**, 22(3): 401-406.

[37] M. Xie, J. Wang, X. Chen, LGH: A Fast and Accurate Algorithm for Single Individual Haplotyping Based on a Two-Locus Linkage Graph. *IEEE/ACM Trans Comput Biol Bioinform*, **2015**, 12(6): 1255-66.

[38] Z.Z. Chen, F. Deng, L. Wang, Exact algorithms for haplotype assembly from whole-genome sequence data. *Bioinformatics*, **2013**, 29(16): 1938-45.

[39] V. Bansal, V. Bafna, HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **2008**, 24(16): i153-9.

[40] J.O. Kitzman, et al., Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol*, **2011**, 29(1): 59-63.

[41] J. Duitama, et al., Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res*, **2012**, 40(5): 2041-53.

[42] E.K. Suk, et al., A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res*, **2011**, 21(10): 1672-85.

[43] H. Cao, et al., De novo assembly of a haplotype-resolved human genome. *Nat Biotechnol*, **2015**, 33(6): 617-22.

[44] S. Levy, The diploid genome sequence of an individual human. *PLoS Biology*, **2007**, 5(10): e254 - e254.

[45] L.M. Li, J.H. Kim, M.S. Waterman, Haplotype reconstruction from SNP alignment. *J Comput Biol*, **2004**, 11(2-3): 505-16.

[46] V. Kuleshov, Probabilistic single-individual haplotyping. *Bioinformatics*, **2014**, 30(17): i379-85.

[47] V. Kuleshov, et al., Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol*, **2014**, 32(3): 261-6.

[48] V. Bansal, et al., An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res*, **2008**, 18(8): 1336-46.

[49] J.H. Kim, M.S. Waterman, L.M. Li, Accuracy assessment of diploid consensus sequences. *IEEE/ACM Trans Comput Biol Bioinform*, **2007**, 4(1): 88-97.

[50] S. Ahn, H. Vikalo, Joint haplotype assembly and genotype calling via sequential Monte Carlo

algorithm. *BMC Bioinformatics*, **2015**, 16: 223.
[51] C.X. Cai, S. Sanghavi, H. Vikalo, Structured Low-Rank Matrix Factorization for Haplotype Assembly. *Ieee Journal of Selected Topics in Signal Processing*, **2016**, 10(4): 647-657.
[52] M. Xie, et al., H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. *Bioinformatics*, **2016**, 32(24): 3735-3744.
[53] E. Berger, et al., HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data. *PLoS Comput Biol*, **2014**, 10(3): e1003502.
[54] D. Aguiar, S. Istrail, Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, **2013**, 29(13): i352-60.
[55] S. Das, H. Vikalo, SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, **2015**, 16: 260.
[56] Z. Puljiz, H. Vikalo, Decoding Genetic Variations: Communications-Inspired Haplotype Assembly. *IEEE/ACM Trans Comput Biol Bioinform*, **2016**, 13(3): 518-30.
[57] J.I. Sohn, J.W. Nam, The present and future of de novo whole-genome assembly. *Brief Bioinform*, **2016**.
[58] K.M. SteinBerg, et al., Building and Improving Reference Genome Assemblies. *Proceedings of the IEEE*, **2017**, 105(3): 422-435.
[59] M.J. Chaisson, R.K. Wilson, E.E. Eichler, Genetic variation and the de novo assembly of human genomes. *Nat Rev Genet*, **2015**, 16(11): 627-40.
[60] E.W. Myers, et al., A whole-genome assembly of Drosophila. *Science*, **2000**, 287(5461): 2196-204.
[61] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, **2008**, 18(5): 821-9.
[62] M.J. Chaisson, P.A. Pevzner, Short read fragment assembly of bacterial genomes. *Genome Res*, **2008**, 18(2): 324-30.
[63] J.T. Simpson, et al., ABySS: a parallel assembler for short read sequence data. *Genome Res*, **2009**, 19(6): 1117-23.
[64] R. Luo, et al., SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **2012**, 1(1): 18.
[65] D. Antipov, et al., hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics*, **2016**, 32(7): 1009-15.
[66] J.A. Reuter, D.V. Spacek, M.P. Snyder, High-throughput sequencing technologies. *Mol Cell*, **2015**, 58(4): 586-97.
[67] I. Garbus, et al., Characterization of repetitive DNA landscape in wheat homeologous group 4 chromosomes. *BMC Genomics*, **2015**, 16: 375.

2. 项目的研究内容、研究目标，以及拟解决的关键科学问题（此部分为重点阐述内容）；

2.1 研究内容

如图 1 所示，本项目在多倍体全基因组测序获得的 DNA 读段数据基础上，研究**无需参考基因组序列的单体型组装方法**，即**多倍体单体型从头组装算法**研究。

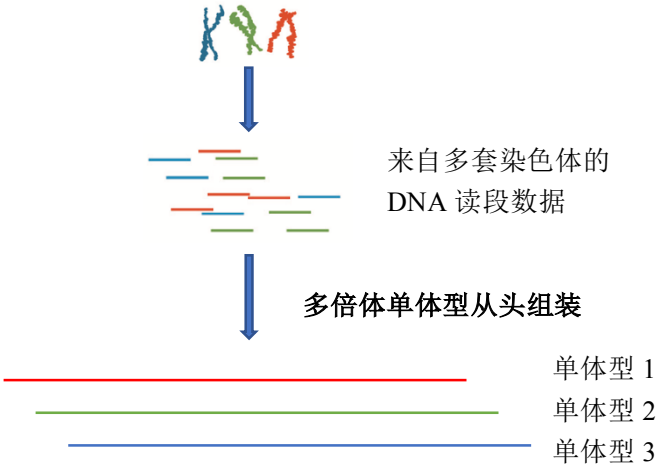

图 1. 研究内容示意图

具体研究内容如下。

（1）异源多倍体基因组序列复杂度分析与测试平台设计

异源多倍体高等植物在进行全基因组测序读段组装时面临巨大的挑战，原因是其**异源多倍性**(allopolyploidy)、**大量重复序列**和**巨大的基因组**。异源六倍体普通小麦（*Triticum aestivum*）是典型的复杂异源多倍体，其体细胞含有 6 套染色体组共 42 条染色体,分别来自 3 个不同的祖先物种，其染色体组成一般用 AABBDD 表示，其基因组包含约 170 亿个碱基对，是人类基因组大小的 5 倍多，其中重复序列在基因组中占 76.6%。2017 年 1 月，普通小麦第一版高质量参考基因组序列[11]已公开发表；异源四倍体陆地棉（*G. hirsutum L.*）的参考基因组序列也已发布[9]。本项目**将对异源四倍体陆地棉、异源六倍体普通小麦的参考基因组序列复杂度进行分析，对重复序列长度分布、异源染色体之间序列差异进行统计分析，挖掘重复序列和异源染色体间序列差异的分布特征，**为后续模型和算法设计及工具的选择提供指导。

为了克服重复序列给单体型组装带来的困难，我们需要把当前不同新一代测序技术平台产生的数据融合起来，使它们优势互补。当前新一代测序平台供应商主要有 Illumina、Roche/454、ABI、Life Technologies、Pacific Biosciences（PacBio）和 Oxford Nanopore Technologies（OxfordNanopores）等。目前被广泛使用的是 Illumina 公司的测序仪，其最新测序仪一次能产生几十亿条长为 150bp（碱基对）

的成对读段（paired-end），其主要测序错误是碱基替换，错误率低于1%[66]，但是其读段长度较短，单纯采用 Illumina 平台生成的 DNA 读段数据，无法组装出包含大量长重复序列的单体型。而 PacBio 的 RS2 测序仪采用单分子实时测序技术，在 4 个小时能产生约 5 万条平均长度大于 14kb 的 DNA 读段，有些读段长度可达 60kb；单分子测序技术生成的读段长，但是测序错误率高达 10%[66]，单纯采用这些数据，也无法有效对复杂多倍体的单体型进行组装。

本项目对融合高通量短读段测序仪和单分子测序仪生成的 DNA 读段数据的多倍体单体型从头组装算法进行研究，因此本项目首先对不同测序技术产生的真实 DNA 读段数据进行统计分析，获得读段数据特征如平均长度、测序误差类型和覆盖深度分布等数据特征，**根据不同测序技术下真实读段数据的典型特征为对应测序技术设计新的或改进已有的模拟数据生成器**，为后续模型和算法研究搭建以异源四倍体陆地棉和异源六倍体普通小麦的参考基因组序列为基准的性能**测试平台**。

（2）高可信 contig 组装与读段数据纠错

在对陆地棉、普通小麦基因组序列复杂度分析以及模拟高通量测序仪、单分子测序仪的数据生成器设计完成以后，本项目将进一步研究**快速的初步组装算法，对高通量测序仪生成的具有较少测序错误的大量短读段进行初步组装，获取高可信 contig**。我们将根据基因组序列复杂度分析结果，选择合适大小的 *k*-mer，利用模拟数据生成器对现有基于 De Bruijn 图的序列组装算法如 Velvet[61]、EULER-SR[62]、ABySS [63]、SOAPdenovo[64]等进行测试比较分析，对其组装出的 contig（De Bruijn 图中一条无分支路径表示的一段连续的 DNA 序列）的准确性进行评估，提出 contig 可信度评价标准，在此基础上，对已有 De Bruijn 图实现方法和组装算法进行改进，设计快速的基于大量短读段的初步组装算法，获取高可信 contig 集合。

然后，对测序仪生成的读段数据中包含的测序错误进行建模，研究**利用高可信 contig 集合把读段中的测序错误与序列变异区分开来、进而进行纠错的方法，对读段，特别是测序错误率高的单分子读段进行纠错**。

（3）基于 contig、paired-end、单分子测序数据和局部组装的多倍体单体型组装优化建模及算法设计

在完成对读段数据纠错之后，本项目将对高可信 contig 及相关的读段采用“重叠-扩展”的方式进行局部组装，获取较长的 DNA 序列，然后利用高通量测序平台的 paired-end 数据、单分子测序平台的长读段数据和局部组装获得的较长 DNA 序列来克服基因组重复序列给单体型组装带来的困扰，研究基于 contig、paired-end、单分子测序长读段数据和局部组装结果的多倍体单体型组装优化模型，并进行算法设计。

项目组将对表示 contig 的 De Bruijn 图进行简化，研究用加权有向图表示 contig、paired-end、单分子测序测序长读段数据和局部组装结果等相关信息的有效方法，根据异源染色体间序列差异分布特性，构建从加权有向图中提取 k-条最优路径以达到 k 个单体型重建目的的优化模型，并设计快速的求解算法。

（4）多倍体从头组装软件包的设计与实现

本项目将在计算机集群上对多倍体单体型从头组装优化模型和算法进行性能测试、分析和优化，对复杂度高、可扩展性低的算法研究其并行及分布式算法，进行软件的设计和开发，为相关生物学家提供多倍单体型从头组装软件包进行测试和分析。

2.2 研究目标

本项目的研究目标是为**融合不同新一代测序数据的多倍体单体型从头组装提供测试平台、计算模型和实用算法支持**。本项目将综合利用不同新一代测序平台生成的 DNA 读段数据，设计有效的计算模型和高效的算法，在缺乏基因组参考序列的条件下，构建以较少的人力和物力确定多倍体各单体型的计算平台。

2.3 拟解决的关键科学问题

（1）**DNA 测序错误与序列变异的区分：** Illumina 平台生成的 DNA 测序数据错误率低于 1%，单分子测序技术生成的读段测序错误率高达 10%，异源六倍体小麦的同源染色体之间的序列差异率约为 0.32%[14]，来自于不同祖先物种的部分同源染色体之间的差异率高达 20%左右[7]，在多倍体单体型从头组装中测序错误和序列变异很难区

分，如何区分 DNA 测序错误与序列变异，对 DNA 读段进行有效的纠错处理，是本项目要解决的关键科学问题之一。

（2）**重复序列邻接的 contig 在单体型中的定位**：复杂多倍体植物大量重复序列是目前阻碍复杂其全基因组测序的主要障碍之一，研究如何利用 paired-end 数据、单分子测序长读段数据和局部组装结果来克服基因组重复序列给单体型从头组装带来的困扰，提出合适的单体型组装优化模型，使重复序列分离的 contig 在单体型中能准确定位是本项目要解决的关键科学问题之二。

（3）**多倍体单体型从头组装优化模型高精度快速求解算法的设计**：复杂问题的有效模型通常是计算难解问题，模型的验证和广泛使用离不开求解的快速算法，根据多倍体各单体型序列差异、重复序列分布等实际生物数据特征，结合参数计算、图论算法、聚类分析等技术为多倍体单体型从头组装优化模型设计出快速高精度算法是本项目要解决的关键科学问题之三。

3. 拟采取的研究方案及可行性分析（包括研究方法、技术路线、实验手段、关键技术等说明）；

3.1. 研究方案

本项目对无需参考基因组序列、利用不同新一代测序平台生成的 DNA 读段数据进行多倍体单体型从头组装的优化模型和算法开展研究。本项目将利用真实和模拟的生物数据对模型的优劣和算法的效率进行验证，对性能较差的模型和算法进行反馈性修正、优化与完善，最终形成多倍体单体型从头组装软件包。研究框架如图 2 所示。

具体研究方案如下：

（1）异源多倍体基因组序列复杂度分析与测试平台设计

异源四倍体陆地棉（*G. hirsutum L.*）的参考基因组序列能通过美国国家生物技术信息中心(National Center of Biotechnology Information，NCBI）的网站（www.ncbi.nlm.nih.gov）和中国农业科学院棉花研究所的棉花基因组工程网站（cgp.genomics.org.cn）免费下载；异源六倍体普通小麦（*Triticum aestivum*）的参考基因组序列也可以通过下面两个网站免费获得：archive.plants.ensembl.org 和 wheat-urgi.versailles.inra.fr。

项目组将下载陆地棉和普通小麦的参考基因组序列，利用已有的

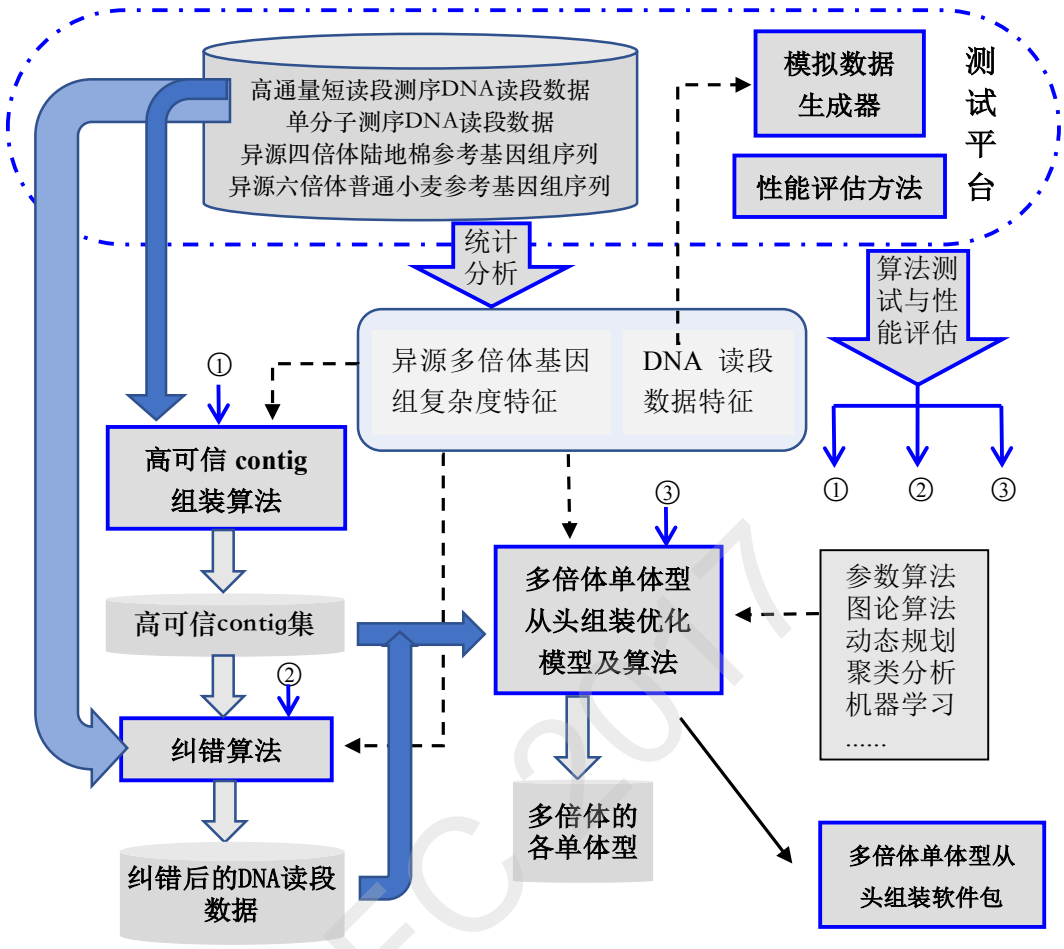

图 2. 研究方案示意图

k-mer 分析工具对其基因组序列中重复序列的大小、位置分布进行统计分析；用已有的同源分析工具比较分析对应的来自不同祖先物种的染色体（部分同源染色体）DNA 序列，对它们之间的相似性与差异性进行统计分析，**获得异源多倍体基因组序列复杂度特征。**

不同测序平台产生的真实 DNA 读段数据可从 NCBI 的数据库 SRA（Sequence Read Archive）中免费下载，本项目将首先选择 Illumina、Pacific Biosciences 这两家公司测序仪生成的一些典型 DNA 读段数据集下载，然后进行统计分析，对不同测序技术生成的 DNA 读段的平均长度、测序误差分布、覆盖深度分布等数据特征进行挖掘，进而**设计出能满足对应数据特征的 DNA 读段模拟数据生成器**，提出对各算法进行性能评价的指标，**搭建以四倍体陆地棉和六倍体普通小麦参考基因组序列为基准的**、能使用真实测序数据和模拟数据**对算法性能进**

行大规模测试的基准平台。

（2）高可信 contig 组装与读段数据纠错

目前有大量基于 De Bruijn 图的序列组装工具，如 Velvet[61]、EULER-SR[62]、ABySS[63]、SOAPdenovo[64]等，项目组根据对基因组序列复杂度分析结果，选择合适大小的 *k*-mer，利用测试平台对较常用的工具在大量的短读段上构建 contig。通过 contig 对基准基因组序列的联配，对其准确性进行评估；分析 contig 的准确性与其在 De Bruijn 图中的拓扑结构、支持该 contig 的读段分布等因素的相关性，提出以 contig 在 De Bruijn 图中的拓扑结构和相关读段覆盖等因素为依据的 contig 可信度评价指标。在此基础上对已有 De Bruijn 图实现方法和 contig 组装算法进行改进，**设计快速的基于短读段的高可信 contig 组装算法**，获取高可信 contig 集合。

为了有效地区分 DNA 测序错误与序列变异，我们将利用高可信 contig 与相关的短读段、单分子测序长读段进行联配，利用机器学习、聚类分析等手段对读段中的测序错误建模，**设计高精度的分类器和纠错算法，根据联配在同一个位点的所有读段的碱基值分布识别 DNA 测序错误，并进行有效的纠错。**

（3）多倍体单体型从头组装优化建模及算法设计

重复序列给单体型组装带来了巨大的挑战。假设一个单体型是“aRbRcRd”，其中 R 是在该单体型中出现了 3 次的重复 DNA 序列，假设其长度为 2kb，a、b、c 和 d 是唯一的 DNA 序列，那么经过高可信 contig 组装后获得的简化 De Buijin 图就会如图 3 所示，如果按照

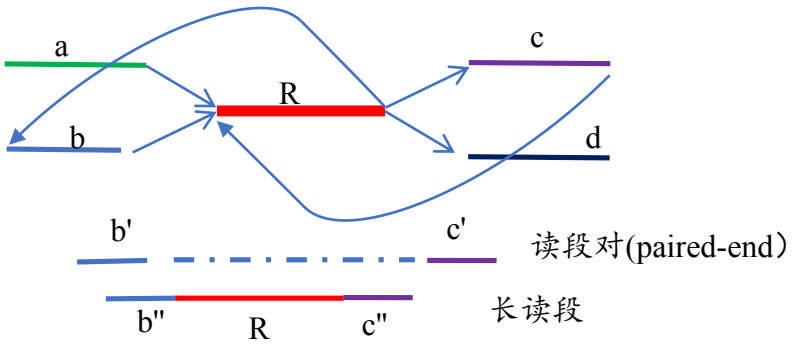

图 3. “aRbRcRd”对应的压缩 De Bruijn 图及相关的读段对和长读段，其中 b'为 b 的子串，c'是 c 的子串，而 b"为 b 的后缀，c"是 c 的前缀。

最长路径优化模型，获得的单体型就有两种可能“aRbRcRd”和“aRcRbRd”。

为了克服**重复序列邻接的 contig 在单体型中的定位困难**，我们将首先利用paired-end的读段成对信息及单分子测序生成的长读段信息。理论上 insert 大小为 *l* kb paired-end 数据或长度为 *l* kb 的长读段能有效解决长度小于 *l* kb 的重复序列给序列组装带来的困难。在图 3 中，依靠 insert 长为 2kb 以上的读段对或长度大于 2kb 的读段就能确定上面压缩 De Bruijn 图对应的单体型应该是“aRbRcRd”而不是“aRcRbRd”。

在异源多倍体单体型组装中，由于来自同源（homologous）染色体和部分同源（homeologous）染色体的 DNA 读段同时存在，重复序列给单体型组装带来的问题更加复杂，本项目还将采用**重叠-扩展**的方式对高可信 contig 及相关读段进行**局部组装**，期望获取长重复序列隔开的 contig 之间的相对位置信息。研究能集成 contig、paired-end 数据、单分子测序数据以及由局部组装结果获得的有效信息的压缩表示形式，计划采用加权有向图，图的节点表示一个 contig，给连接节点的有向边赋予一系列权值（d, n），表示支持这两个 contig 位于同一单体型上相距 d 个碱基的读段数为 n。根据异源多倍体复杂度特征，寻求合理的优化模型，通过构建加权有向图中的 *k*-条最优路径来确定**重复序列邻接 contig 在单体型中的最优定位。**

我们计划将二倍体单体型组装中的 MEC 模型移植到多倍体单体型从头组装问题中，初步的构想是在有向加权图中寻找 k 条路径，能覆盖最多的 contig 且不兼容边的权重之和最小。这样的优化模型可能是难解的计算问题，为了设计快速精确的求解算法，**项目组将把参数计算、动态规划与高级图论算法等技术结合起来进行精确算法和启发式算法设计。**我们将利用本项目组第一步搭建的平台对提出的模型和算法进行大量的测试，对它们的性能指标进行大量对比分析，发现各模型和算法的优缺点。进而通过对模型内在结构及生物学意义进行深入挖掘，分析模型依据的生物意义与其单体型重建精度之间的关系，反复“**模型提出→模型性能测试和分析→模型修改**”这一过程，**最终提出融合高通量短读段数据和单分子测序长读段数据的高精度多倍**

体单体型从头组装优化模型，并设计出快速有效的算法。

3.2 可行性分析

NCBI 的 SRA 数据库提供包括 Illumina, 454, IonTorrent, Complete Genomics, PacBio 和 OxfordNanopores 等测序平台的真实测序读段数据下载，异源六倍体小麦和四倍体棉花的参考基因组序列也已公开发布，这些数据源给本项目所需的真实生物数据提供了可靠的保证。

虽然多倍体植物基因组重复序列在整个基因组中的比例非常高，如普通小麦中占 76.6%，可是绝大部分单个重复序列长度较短，Garbus 等[67]对普通小麦 3 个部分同源的 4 号染色体 4A，4B 和 4D 的重复序列进行分析后发现，反转录因子(Retroelement)占重复序列的绝大部分，其平均长度小于 1kb，而 PacBio 的单分子测序平均读段长度为 14kb，理论上是可以克服普通小麦绝大部分重复序列给单体型从头组装带来的困难。因此**融合高通量短读段测序技术的高测序精度和单分子测序平台的长读段对多倍体单体型进行从头组装，从理论上是可行的**，并且随着测序技术的不断发展，测序价格快速下降而测序的精度、读段的长度不断提高，多倍体单体型从头组装将成为获得多倍体植物基因组序列最经济的方式。

基因组重复序列分析和异源多倍体的部分同源染色体序列比较均可依赖成熟的生物信息学工具，而“模拟数据生成器设计及测试平台搭建”申请人在前一个已完成的国家自然科学面上项目中有类似的经验和成果，研究内容的第（1）项是可行的。

利用 De Bruijn 图对大量来自巨大基因组的短读段进行 contig 组装，可以有效节省内存，有大量有效的算法可供借鉴，通过大量模拟测试，提出以 contig 在 De Bruijn 图中的相关拓扑结构和相关读段覆盖为依据的 contig 可信度评价指标，进而设计高可信 contig 组装算法，在研究方法上是可行的。利用高可信 contig 及相关读段的联配，获得联配位点上 DNA 取值的大量分布数据很适合利用机器学习、聚类分析等方法构造分类器对测序错误进行识别。研究内容的第（2）项是可行的。

利用高可信 contig，采用 contig 与相关读段采用重叠-扩展的方式进行局部序列组装，能以较少的计算资源构建一个较长连续 DNA

序列，结合单分子测序的长读段和相距一定距离的成对读段，计算重复序列邻接的 contig 在单体型中的拓扑顺序，是克服重复序列给单体型组装带来困扰的有效方法。利用有向加权图集成上述信息，能有效降低后续计算对内存的巨大需求。通过优化建模，寻找不兼容信息最少的 k 条路径来重建多倍体单体型符合经纠错后读段错误率很小的实际情况。经纠错后，DNA 读段的错误率有可能降到远低于 1%，同源染色体序列的片段差异也不是很大，一般为 0.3%，而通过对有向加权图的拓扑特征进行统计分析，有可能发现如节点的入度和出度均较少的小参数特性，项目组研究成员对参数计算、图算法、动态规划、聚类分析、机器学习都有深刻的理解，并有成功的应用（参照研究基础部分和主要参与人简历），充分利用这些小参数特性，把参数计算、图算法、动态规划、聚类分析、机器学习等方法结合起来设计快速的精确算法或启发式算法在研究方法上是可行的。据此，研究内容的第（3）项也是可行的。

从实施条件看，本项目所依托的单位具有丰富的计算资源，能提供高性能计算机集群作为本项目的计算平台。从项目组成员的科研能力看，申请人在前一个国家自然科学基金的资助下，在**有参考基因组序列的二倍体单体型组装和多倍体单体型组装问题的建模和算法**上有深入的研究，取得了国际领先的研究成果，相关研究成果在 Bioinformatics 等生物信息学知名刊物上发表，“面向生物特征的数据处理理论与方法” 2013 年获教育部自然科学二等奖（申请人排名第 3）。项目组主要研究成员姜涛教授是生物信息学和计算机算法领域国际知名专家，在转录组序列组装、图算法、聚类算法等方面的学术论文被同行专家大量引用。项目组其他研究人员在生物信息学、计算机算法等领域进行了较长时间的研究工作，具有较强的科研能力，可以说，本项目组已经积累了比较丰富的研究经验，打下了比较坚实的研究基础，在本项目的资助下有能力取得更突出的研究成果。

总之，本项目的应用需求强烈，研究目标明确，研究内容具体，反映了目前生物信息学研究领域的热点问题。本项目的研究方法、技术路线及实验方案切实可行，在国内外具有先进性。本项目已充分具备理论、技术和条件上的可行性。

4. 本项目的特色与创新之处；

（1）无需基因组参考序列，融合高通量短读段数据和单分子测序长读段数据对多倍体单体型进行从头组装。

（2）提出基于拓扑结构和相关读段覆盖为依据的 contig 的可信度评价指标，利用高可信 contig 对读段纠错。

（3）利用加权有向图集成 contig、局部组装结果、paired-end 和长读段数据相关信息，把多倍体单体型组装转化为加权有向图上的优化问题。

5. 年度研究计划及预期研究结果（包括拟组织的重要学术交流活动、国际合作与交流计划等）。

5.1 年度研究计划

（1）2018 年 1 月至 2018 年 12 月：完成异源多倍体基因组序列复杂度分析与测试平台的设计；提出基于拓扑结构和相关读段覆盖为依据的 contig 可信度评价指标，设计基于 De Bruijn 图的高可信 contig 组装算法。

（2）2019 年 1 月至 2019 年 12 月：对测序错误和序列多态性进行建模，设计基于高可信 contig 的纠错算法；研究基于高可信 contig 的局部组装算法；设计加权有向图的压缩存储结构以集成 contig、局部组装结果、paired-end 和长读段数据，对基于加权有向图的多倍体单体型从头组装优化模型进行初步研究。

（3）2020 年 1 月至 2020 年 12 月：提出基于加权有向图的多倍体单体型从头组装优化模型，利用参数计算、图算法、动态规划、机器学习、聚类分析等方法进行算法设计。

（4）2021 年 1 月至 2021 年 12 月：继续研究单体型从头组装优化模型及算法，对已提出的模型和算法进行修改和完善；多倍体单体型从头组装软件包的设计与实现；资料整理、研究成果汇报和项目结题。

5.2 预期研究结果

本项目以已经发布的复杂多倍体植物基因组参考序列、高通量短读段测序平台及单分子测序平台产生的大量真实生物数据为基础，研究无需基因组参考序列，融合高通量短读段测序及单分子测序数据的

多倍体单体型从头组装优化模型和算法。预期研究结果包括：

（1）设计出基于 De Bruijn 图的**高可信 contig 组装算法**，及基于高可信 contig 的**纠错算法**；

（2）提出符合实际测序数据特征及多倍体基因组复杂度特征的基于加权有向图的**多倍体单体型从头组装优化模型**；

（3）设计出求解**多倍体单体型从头组装优化模型**的**高精度快速算法**。

以上预期研究结果将具体表现为：

（1）在 3 次以上的知名国际会议上宣读研究成果，在国际权威学术期刊上发表 8 篇以上高水平研究论文，在国际学术会议和国内核心期刊上发表的相关研究论文 10 篇左右；

（2）申请多倍体单体型从头组装国家发明专利 1-2 项；

（3）设计出实用的多倍体单体型从头组装软件包，申请相关软件著作权 2-3 项；

（4）培养在生物信息学领域具有较高研究水平的博士研究生 4 名左右，硕士研究生 6 名左右。

国际合作与交流计划安排情况

（1）参加国际会议 4-8 人次；

（2）加强国际合作，邀请国外生物信息学领域知名专家 1-2 名进行短期合作研究；

（3）邀请生物信息学及算法设计领域的国外知名学者进行 3 次以上讲学和讨论。

（二）研究基础与工作条件

1. 研究基础（与本项目相关的研究工作积累和已取得的研究工作成绩）；

项目组主要研究人员近几年来在生物信息学、参数计算理论、图论算法、机器学习和聚类分析等领域进行了深入的研究和实践，积累了相当多的研究成果和经验。

申请人曾承担国家自然科学基金面上项目“**新一代测序技术下单体型组装问题计算模型和算法研究**”，对**有基因组参考序列的二倍体单体型组装**问题进行了深入研究，取得了国际领先的研究成果。申请

人成功地运用参数计算理论于二倍体单体型组装问题，提出的参数化精确算法显著降低了求解单体型组装计算模型 MSR、MFR 及其他计算模型的时空复杂度，相关的研究成果发表在**计算机学报**、**Algorithmica** 和多个知名国际会议上。申请人对二倍单体型组装提出了三个新的优化模型并设计了相应的快速求解算法：第一个是基于有误差的基因型和片段数据的计算模型 WMLF/GS, 研究论文被生物信息学领域国际顶级会议 The 16th Annual International Conference Intelligent Systems for Molecular Biology（**ISMB 2008**, Toronto, Canada）接受，全文发表在生物信息学国际知名杂志 **Bioinformatics** 上[26]；第二个为平衡分区模型（BOP），相关成果发表在 **BMC Systems Biology** 上[22]；第三个把 MEC 转化为位点连锁图最优标签模型，相关成果发表在 IEEE/ACM Transactions on Computational Biology and Bioinformatics（**TCBB**）上[37]。项目相关研究成果“面向生物特征的数据处理理论与方法”于 2013 年获**教育部自然科学二等奖**（申请人排名第 3）。

申请人对**有基因组参考序列的多倍体单体型组装**问题也进行了初步研究。申请人把多倍体单体型组装建模成 DNA 读段的多分区优化模型，通过限制动态规划迭代过程中中间解的个数，设计了能平衡求解精度和求解速度的两个算法 H-PoP 和 H-PoPG，大量实验测试结果表明 H-PoP 和 H-PoPG 的性能要明显优于已有的多倍体单体型组装算法 HapTree[53]、HapCompass[54]和 SDhaP[55]，相关研究成果发表在 **Bioinformatics** 上[52]。

项目组主要研究成员姜涛教授，在生物信息学、计算机算法等领域研究成果众多。姜涛教授在基于家族的单体型推断问题取得了国际领先的研究成果，在 ZRHC、MRHC 和 *k*-RHC 等计算模型的计算复杂度证明、精确或近似算法设计等方面发表了一系列高水平学术论文，开发的相关软件工具 PedPhase 被大量下载使用。姜涛教授在“Heilbronn Triangle”问题上取得的成果曾被 New Scientist、Math Digest、Courrier International 等知名科普杂志和报纸广为报道；在转录组序列组装、最短公共超串（shortest common superstrings）、有向 Steiner 树、寡核苷酸指纹聚类（Clustering of oligonucleotide fingerprints）等问题的算法设计上也取得了大量高水平的研究成果。姜涛教授近年来参与

了加州大学河滨分校的**大麦（Barley）基因组测序工程**（见参与者简介中（一、3（2））发表在 The Plant Journal, 2015 上的论文），其在有大量重复序列的大麦基因组组装上的方法和经验将为本项目关键科学问题的解决和快速算法设计提高可靠的理论指导。

项目组主要成员钟坚成副教授，利用聚类分析、机器学习等手段对蛋白质交互网络进行分析，基于海量数据对关键蛋白质进行预测，其研究成果在 TCBB 等国际知名期刊上发表。钟坚成在海量数据处理、聚类分析、机器学习等方面的经验可用于本项目的算法设计。

综上所述，本项目组在复杂问题建模、参数化算法、图论算法、动态规划、聚类分析、机器学习等方面有良好的研究基础，在相关领域取得了国际领先的研究成果，本项目具有良好的研究基础。

2．工作条件（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门重点实验室等研究基地的计划与落实情况）；

项目申请人所在单位湖南师范大学物理与信息科学学院拥有很好的实验环境。湖南师范大学物理与信息科学学院，在“211”工程和“低维量子结构与调控”教育部重点实验室的连续资助下，拥有良好的研究实验条件、齐全的软硬件设备，拥有价值几百万的计算设备。可供本项目组使用的计算资源主要有拥有 16 个计算节点的计算机集群，该集群每个计算节点拥有 8 个高性能 4 核 CPU,有两个节点拥有 256GB 内存,其他 14 个计算节点拥有 64GB 内存，集群拥有 3.75TB 的存储设备。另外湖南师范大学高性能计算与随机信息处理省部共建教育部重点实验室也能为本项目提供计算资源，该重点实验室拥有基于曙光 5000A 和 TESLA 架构的万亿次高性能并行计算机和一套曙光天潮高性能计算机集群系统。这些高性能设备为本项目进行资源密集的生物问题计算提供了良好的支持平台。

3．正在承担的与本项目相关的科研项目情况（申请人和项目组主要参与者正在承担的与本项目相关的科研项目情况，包括国家自然科学基金的项目和国家其他科技计划项目，要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等）；

无

4. 完成国家自然科学基金项目情况（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限 500 字）和相关成果的详细目录）。

申请人在 2011 年--2013 年期间承担了国家自然科学基金面上项目“**新一代测序技术下单体型组装问题计算模型和算法研究**”（批准号：61070145）的研究，项目组主要研究了新一代测序技术下**二倍体单体型组装**计算问题组合优化模型构建及高效算法设计。研究基本按照计划执行，较好地完成了预期任务。项目组主要利用参数计算理论，建立固定参数优化模型，设计高效参数化算法，在 Bioinformatics、BMC Systems Biology、BMC Bioinformatics 等国际知名刊物上发表学术论文 5 篇，在小型微型计算机系统等国内核心刊物发表论文 3 篇；培养硕士研究生 3 名。

项目结题后，项目组对单体型组装问题进行了继续研究，提出了一个新的两倍体单体型重建算法 LGH，研究成果于 2015 年发表在国际知名期刊 IEEE/ACM Transactions on Computational Biology and Bioinformatics（TCBB）上。同时，项目组对多倍体单体型组装问题进行了初步研究，设计了相关算法 H-PoP 和 H-PoPG，研究成果于 2016 年发表在国际知名期刊 Bioinformatics 上。项目组在单体型组装问题上的研究成果为**本申请项目“多倍体单体型从头组装算法研究”**提供了较好的研究基础，但**多倍体单体型从头组装问题的计算复杂性远大于二倍体单体型组装**，本申请项目是该项目的自然扩展。

附总结摘要

在基金的支持下，课题组在新一代测序技术下单体型组装计算问题组合优化模型的构建、高效算法的设计与分析上取得了显著的进展。通过对单体型组装问题相关真实生物数据的整合和数据特征的抽取，课题组完成了模拟数据生成器的设计和测试平台的建设;结合图论和聚类分析等技术提出了新的单体型组装模型;根据单体型组装实质就是利用读段的两个杂合 SNP 位点之间的组合模式来组装一条染色体上较大区域的 SNP 序列，把片段数据转化成两位点连锁图，提出了

新的连锁图标签优化单体型组装模型。通过对新一代测序技术下真实生物数据特征的挖掘，课题组发现了输入数据的一些小参数特征，进而利用参数计算理论，为多个优化模型设计了快速的参数化动态规划精确算法。动态规划递推过程中要保留的中间结果的多少是影响算法时空复杂度的决定因素，课题组发现只保留一部分较优的中间结果能加快算法的速度，而对最终性能没有显著影响，据此，课题组进一步设计了基于 top-k 个中间最优解的启发式动态规划算法。项目促进了单体型计算模型及算法研究，项目的研究成果为生物信息学中大量复杂计算问题实用算法设计提供了新思路，也将促进单体型在复杂疾病全基因关联分析中的应用。

相关成果

论文：

(1) Minzhu Xie, Qiong Wu, Jianxin Wang, and Tao Jiang. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics, 2016, 32(24):3735-3744.

(2) Minzhu Xie, Jianxin Wang, and Xin Chen. LGH: A Fast and Accurate Algorithm for Single Individual Haplotyping Based on a Two-Locus Linkage Graph. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12(6):1255-66.

(3) Minzhu Xie, and Jing Wang. Parameterised algorithms of the individual haplotyping problem with gaps. Int. J. Bioinformatics Research and Applications，2013, 9(1):25-40.

(4) Yang Yang, and Xie Minzhu. A parameterized enumeration algorithm for the haplotype assembly problem. in Proc. of 6th International Conference on Biomedical Engineering and Informatics (BMEI), 2013, pp. 458-462.

(5) Minzhu Xie, Jianxin Wang, and Tao Jiang. A fast and accurate algorithm for single individual haplotyping. BMC Systems Biology, 2012, 6(Suppl 2):S8.

(6) Minzhu Xie, Jing Li, and Tao Jiang. Detecting genome-wide epistases based on the clustering of relatively frequent items. Bioinformatics, 2012, 28(1):5-12.

(7) Minzhu Xie, Jianxin Wang, and Jianer Chen. A practical parameterised algorithm for the individual haplotyping problem MLF. Mathematical Structures in Computer Science, 2010, 20(5):851-863.

(8) Minzhu Xie, Jing Li, and Tao Jiang. Accurate HLA type inference using a weighted similarity graph. BMC Bioinformatics, 2010, 11(Suppl 11):S10.

(9) 谢民主, 罗锋, 唐烽. 单体型组装最大片段割参数化精确算法. 小型微型计算机系统, 2014, 35(2):353 - 357.

(10) 谢民主, 刘新求, 杨洋. 两位点疾病模型的快速参数求解算法. 计算机工程, 2012, 38(19):266-268, 273.

(11) 谢民主, 杨洋. 复杂疾病模型快速参数求解算法. 计算机工程与应用, 2012, 48(7):121-123.

奖励：

(1) 2013 年教育部自然科学二等奖: 面向生物特征的数据处理理论与方法。主要完成人：王建新，李敏，谢民主，冯启龙。主要完成单位：中南大学，湖南师范大学。

(2) 2012 年湖南省第 14 届自然科学二等优秀学术论文: A practical parameterised algorithm for the individual haplotyping problem MLF, 谢民主, 王建新, 陈建二。

（三）其他需要说明的问题

1. 申请人同年申请不同类型的国家自然科学基金项目情况（列明同年申请的其他项目的项目类型、项目名称信息，并说明与本项目之间的区别与联系）。

无

2. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在同年申请或者参与申请国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，申请或参与申请的其他项目的项目类型、项目名称、单位名称、上述人员在该项目中是申请人还是参与者，并说明单位不一致原因。

无

3. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在与正在承担的国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，正在承担项目的批准号、项

目类型、项目名称、单位名称、起止年月，并说明单位不一致原因。

无

4. 其他。

无

谢民主 简历

湖南师范大学，物理与信息科学学院，教授

教育经历（从大学本科开始，按时间倒序排序；请列出攻读研究生学位阶段导师姓名）：

1. 2004/9－2008/5， 中南大学， 计算机应用技术， 博士， 导师：陈建二，王建新
2. 2000/9－2003/6， 中南大学， 计算机应用技术， 硕士， 导师：蒋外文，王加阳
3. 1987/9－1990/6， 湖南教育学院， 物理学， ， 导师：
4. 1983/9－1986/6， 娄底中等师范学校， 中小学教育， ， 导师：

科研与学术工作经历（按时间倒序排序；如为在站博士后研究人员或曾进入博士后流动站（或工作站）从事研究，请列出合作导师姓名）：

1. 2013/12-至今，湖南师范大学，物理与信息科学学院，教授
2. 2010/1-2011/5，美国加州大学河滨分校，计算机系，博士后
3. 2008/11-2013/11，湖南师范大学，物理与信息科学学院，副教授
4. 2003/7-2008/10，湖南师范大学，物理与信息科学学院，讲师
5. 1986/7-2000/8，湖南涟源教育局，白马镇中学，中学教师
6. 2008/12-2011/12，中南大学，博士后，合作导师：桂卫华，陈松乔，王建新

曾使用其他证件信息（申请人应使用唯一身份证件申请项目，曾经使用其他身份证件作为申请人或主要参与者获得过项目资助的，应当在此列明）：

主持或参加科研项目（课题）及人才计划项目情况：

1. 国家自然科学基金面上项目，61370172，全基因组关联分析基因交互作用探测算法研究，2014/01-2017/12，73万元，在研，主持
2. 国家自然科学基金面上项目，61070145，新一代测序技术下单体型组装问题计算模型和算法研究，2011/01-2013/12，32万元，已结题，主持
3. 国家自然科学基金面上项目，60773111，参数计算理论及应用，国家自然科学基金，2008/01-2010/12，30万，已结题，参与(排名第2)

代表性研究成果和学术奖励情况

（请注意：①投稿阶段的论文不要列出；②对期刊论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、期刊名称、发表年代、卷（期）及起止页码（摘要论文请加以说明）；③对会议论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、会议名称(或会议论文集名称及起止页码)、会议地址、会议时间；④应在论文作者姓名后注明第一/通讯作者情况：所有共同第一作者均加注上标“#”字样，通讯作者及共同通讯作者均加注上标“*”字样，唯一第一作者且非通讯作者无需加注；⑤所有代表性研究成果和学术奖励中本人姓名加粗显示。）

一、期刊论文

1. 第一作者论文

(1) **Minzhu Xie**, Qiong Wu, Jianxin Wang, Tao Jiang, H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids, Bioinformatics, 2016, 32 (24) : 3735~3744

(2) **Minzhu Xie**(*), Jianxin Wang, Xin Chen(*), LGH: a fast and accurate algorithm for single individual haplotyping based on a two-locus linkage graph, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12 (6) : 1255~1266

(3) **谢民主**，罗锋，唐烽，单体型组装最大片段割参数化精确算法，小型微型计算机系统，2014，35（2）：353~357

(4) **Minzhu Xie**, Jing Wang, Parameterised algorithms of the individual haplotyping problem with gaps, International Journal of Bioinformatics Research and Applications, 2013, 9 (1) : 25~40

(5) **Minzhu Xie**, Jianxin Wang, Tao Jiang, A fast and accurate algorithm for single individual haplotyping, BMC Systems Biology, 2012, 6 (Suppl 2) : S8~S8

(6) **Minzhu Xie**(*), Jing Li, Tao Jiang(*), Detecting genome-wide epistases based on the clustering of relatively frequent items, Bioinformatics, 2012, 28 (1) : 5~12

(7) **Minzhu Xie**(*), Jing Li, Tao Jiang(*), Accurate HLA type inference using a weighted similarity graph, BMC Bioinformatics, 2010, 11 (S11) : S10

(8) **Minzhu Xie**(*), Jianxin Wang(*), Jianer Chen, A practical parameterised algorithm for the individual haplotyping problem MLF, Mathematical Structures in Computer Science, 2010, 20 (5) : 851~863

(9) **Minzhu Xie**, Jianxin Wang(*), Jianer Chen, Jingli Wu, Xucong Liu, Computational Models and Algorithms for the Single Individual Haplotyping Problem, Current Bioinformatics, 2010, 5 (1) : 18~28

(10) **谢民主**，陈建二，王建新，个体单体型问题参数化算法研究，计算机学报，2009，32（8）：1637~1650

(11) **Minzhu Xie**, Jianxin Wang, An improved (and practical) parameterized algorithm for the individual haplotyping problem MFR with mate-pairs, Algorithmica, 2008, 52 (2) : 250~266

(12) **Minzhu Xie**, Jianxin Wang(*), Jianer Chen, A model of higher accuracy for the individual haplotyping problem based on weighted SNP fragments and genotype with errors, Bioinformatics (also presented at ISMB 2008, Toronto, July 19 2008), 2008, 24 (13) : I105~I113

(13) **谢民主**，陈建二，王建新，有Mate-Pairs的个体单体型MSR问题的参数化算法，软件学报，2007，18（9）：2070~2082

2. 通讯作者论文（勿与第一作者论文重复）

(1) Jianxin Wang，**Minzhu Xie**(*)，Jianer Chen，A Practical Exact Algorithm for the Individual Haplotyping Problem MEC/GI，Algorithmica，2010，56（3）：283~296

3. 既非第一作者又非通讯作者论文

(1) Xiaojun Ding，Jianxin Wang，A. Zelikovsky，X. Guo，**Minzhu Xie**，Yi Pan，Searching high-order SNP combinations for complex diseases based on energy distribution difference，IEEE/ACM Transactions on Computational Biology and Bioinformatics，2015，12（3）：695~704

二、获得学术奖励

(1) **谢民主**（3/4），面向生物特征的数据处理理论与方法，中华人民共和国教育部，自然科学，省部二等奖，2014.1.29
（王建新，李敏，**谢民主**，冯启龙）

除非特殊说明，请勿删除或改动简历模板中蓝色字体的标题及相应说明文字

参与者 简历

姜涛，清华大学，清华大学信息科学与技术国家实验室(筹)，千人讲座教授

教育经历（从大学本科开始，按时间倒序排序；请列出攻读研究生学位阶段导师姓名）：

1985/01-1988/11，美国明尼苏达大学，计算机系，博士，导师：Oscar H. Ibarra

1979/09-1984/07，中国科技大学，计算机系，学士

科研与学术工作经历（按时间倒序排序；如为在站博士后研究人员或曾进入博士后流动站（或工作站）从事研究，请列出合作导师姓名）：

1. 2014/06-至今，清华大学, 信息科学与技术国家实验室（筹）, 千人讲座教授
2. 1999/07-至今，美国加州大学河滨分校, 计算机系, 教授
3. 2008/05-2014/05，清华大学, 信息科学与技术国家实验室（筹）, Michael Waterman 访问讲席教授
4. 2007/7-2010/6，美国加州大学河滨分校, 计算机系, 校长讲座教授
5. 2006/02-2009/01，清华大学, 计算机系, 长江学者讲座教授
6. 2003/08-2006/07，清华大学, 计算机系, 姚期智访问讲席教授
7. 2002/09-2005/09，北京大学, 理论生物中心, 访问教授
8. 1998/07-2001/06，加拿大McMaster大学, 计算与软件系, 教授
9. 1993/07-1998/07，加拿大McMaster大学, 计算与软件系, 副教授
10. 1989/01-1993/07，加拿大McMaster大学, 计算与软件系, 助理教授

曾使用其他证件信息（申请人应使用唯一身份证件申请项目，曾经使用其他身份证件作为申请人或主要参与者获得过项目资助的，应当在此列明）

护照，BA422887

主持或参加科研项目（课题）及人才计划项目情况（按时间倒序排序）：

1. 国家自然科学基金青年项目，61502027，更高效的PacBio长read纠错算法的研究，2016/01-2018/12，在研，参加
2. 中组部千人计划（创新短期），2014年
3. 国家自然科学基金面上项目，61370172，全基因组关联分析基因交互作用探测算法研究，2014/01-2017/12，72万元，在研，参加
4. 国家自然科学基金面上项目，61175002，融合多种表型相似性和基因相似

性的疾病关联基因预测方法，2012/01-2015/12，59万元，已结题，参加

5. 国家自然科学基金杰青B类，60528001，关于基因网络的算法研究，40万元，2006/01-2008/12，已结题，主持

6. 教育部长江学者（讲座教授），2006年-2009年，计算机科学与技术

代表性研究成果和学术奖励情况（每项均按时间倒序排序）

（请注意：①投稿阶段的论文不要列出；②对期刊论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、期刊名称、发表年代、卷（期）及起止页码（摘要论文请加以说明）；③对会议论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、会议名称(或会议论文集名称及起止页码)、会议地址、会议时间；④应在论文作者姓名后注明第一/通讯作者情况：所有共同第一作者均加注上标“#”字样，通讯作者及共同通讯作者均加注上标“*”字样，唯一第一作者且非通讯作者无需加注；⑤所有代表性研究成果和学术奖励中本人姓名加粗显示。）

一、期刊论文（仅不列此项时可删除该标题）

1. 第一作者论文（仅不列此项时可删除该标题）

(1) **T. Jiang**, P. Kearney, M. Li, A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application, SIAM Journal on Computing, 2001, 30(6): 1942-1961

(2) **T. Jiang**, M. Li, P. Vitanyi, A lower bound on the average-case complexity of Shellsort, Journal of the ACM, 2000, 47(5): 905–911

(3) **T. Jiang**, P. Kearney, M. Li, Some open problems in computational molecular biology. Journal of Algorithms, 2000, 34(1): 194-201 (invited)

(4) **T. Jiang,** M. Li, P. Vitanyi, New applications of the incompressibility method, Computer Journal, 1999, 42(4): 287-293 (invited)

(5) **T. Jiang**, J. Seiferas, P. Vitanyi, Two heads are better than two tapes, Journal of the ACM, 1997, 44(2): 237-256

(6) **T. Jiang**, M. Li, DNA sequencing and string learning. Mathematical Systems Theory, 1996, 29(4): 387- 405

2. 通讯作者论文（勿与第一作者论文重复）（仅不列此项时可删除该标题，序号按实际情况编排）

(1) E. Yang*, **T. Jiang***, SDEAP: a splice graph based differential transcript

expression analysis tool for population data, Bioinformatics, 2016, 32(23): 3593-3602

(2) J. Liu, T. Yu, **T. Jiang***, G. Li*, TransComb: Genome-guided transcriptome assembly via combing junctions in splicing graphs, Genome Biology 2016, 17:213, DOI: 10.1186/s13059-016-1074-1

(3) S. Ma, **T. Jiang***, R. Jiang*, Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data, Bioinformatics, 2015, 31(4): 563-571

(4) M. Tasnim, S. Ma., E.W. Yang, **T. Jiang***, W. Li*, Accurate inference of isoforms from multiple sample RNA-Seq data, BMC genomics, 2015, 16(Suppl 2): S15 (also presented at the 13th Asia Pacific Bioinformatics Conference (APBC), Hsinchu, Taiwan, Jan., 2015 (the best paper award))

(5) O. Tanaseichuk*, J. Borneman, **T. Jiang***, Phylogeny-based classification of microbial communities, Bioinformatics, 2014, 30(4):449-456

(6) E.W. Yang*, T. Girke, **T. Jiang***, Differential gene expression analysis using coexpression and RNA-Seq data, Bioinformatics, 2013, 29(17): 2153-2161

(7) W. Li, **T. Jiang***, Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads, Bioinformatics, 2012: 28(22):2914-2921

3. 既非第一作者又非通讯作者论文（仅不列此项时可删除该标题，序号按实际情况编排）

(1) M. Xie, Q. Wu, J. Wang, **T. Jiang**, H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids, Bioinformatics, 2016, 32(24): 3735-3744

(2) M. Munoz-Amatriain[#], S. Lonardi[#], M. Luo, K. Madishetty, J. T. Svensson, M. J. Moscou, S. Wanamaker, **T. Jiang**, A. Kleinhofs, G. J. Muehlbauer, R. P. Wise, N. Stein, Y. Ma, E. Rodriguez, D. Kudrna, P. R. Bhat, S. Chao, P. Condamine, S. Heinen, J. Resnik, R. Wing, H. N. Witt, M. Alpert, M. Beccuti, S. Bozdag, F. Cordero, H. Mirebrahim, R. Ounit, Y. Wu, F. You, J. Zheng, H. Simkova, J. Dolezel, J. Grimwood, J. Schmutz, D. Duma, L. Altschmied, T. Blake, P. Bregitzer, L. Cooper, M. Dilbirligi, A. Falk, L. Feiz, A. Graner, P. Gustafson, P. M. Hayes, P. Lemaux, J. Mammadov, T. J. Close*, Sequencing of 15622 gene-bearing BACs clarifies the gene-dense regions of the barley genome, The Plant Journal, 2015, 84(1): 216-227

(3) C. Xu, X. Ju, D. Song, F. Huang, D. Tang, Z. Zou, C. Zhang, T. Joshi, L. Jia, W. Xu, K.-F. Xu, Q. Wang, Y. Xiong, Z. Guo, X. Chen, F. Huang, J. Xu, Y. Zhong, Y. Zhu, Y. Peng, L. Wang, X. Zhang, R. Jiang, D. Li**, T. Jiang**, D. Xu, C. Jiang, An association analysis between psychophysical characteristics and genome-wide gene expression changes in human adaptation to the extreme climate at the Antarctic Dome Argus, Molecular Psychiatry, 2015, 20(4):536-544

(4) E. Bao, **T. Jiang**, T. Girke*, AlignGraph: algorithm for secondary *de novo* genome assembly guided by closely related references, Bioinformatics, 2014, 30(12): i319-i328 (also presented at the 22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), July, 2014, Boston, MA)

(5) E. Bao, **T. Jiang**, T. Girke*, BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences, Bioinformatics, 2013, 29(10):1250-1259

(6) B. Fang, D. Mane-Padros, E. Bolotin, **T. Jiang**, F. Sladek*, Identification of a binding motif specific to HNF4 by comparative analysis of multiple nuclear receptors, Nucleic Acids Research, 2012, 40(12):5343-5356

(7) Y. Pirola*, P. Bonnizoni, **T. Jiang**, An efficient algorithm for haplotype inference on pedigrees with recombinations and mutations, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(1): 12-25

二、会议论文（仅不列此项时可删除该标题，标题序号按实际情况编排）

1. 通讯作者论文（勿与第一作者论文重复）（仅不列此项时可删除该标题，序号按实际情况编排）

(1) E. Yang*, **T. Jiang***, GDNorm: An improved Poisson regression model for reducing biases in Hi-C data. 14th Workshop on Algorithms in Bioinformatics, WABI 2014, Wroclaw, Poland, 2014.9.8-9.10

(2) O. Tanaseichuk*, J. Borneman, **T. Jiang***, A probabilistic approach to accurate abundance-based binning of metagenomics reads, 12th Workshop on Algorithms in Bioinformatics, WABI 2012, Ljubljana, Slovenia, 2012.9.10-9.12

(3) O. Tanaseichuk*, J. Borneman, **T. Jiang***, Separating metagenomic short reads into genomes via clustering (extended abstract), 11th Workshop on Algorithms in Bioinformatics, WABI 2011, Saarbruken, Germany, 2011.9.5-9.7

(4) W. Li*, J. Feng, **T. Jiang***, IsoLasso: A LASSO regression approach to

RNA-Seq based transcriptome assembly (extended abstract), 15th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2011, Vancouver, BC, Canada, 2011.3.28-3.31

(5) J. Feng*, W. Li, **T. Jiang***, Inference of isoforms from short sequence reads (extended abstract), 14th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2010, Lisbon, Portugal, 2010.8.12-15

2. 既非第一作者又非通讯作者论文（仅不列此项时可删除该标题，序号按实际情况编排）

(1) S. Zhang, H. Hu, J. Zhou, X. He, **T. Jiang**, J. Zeng, ROSE: a deep learning based framework for predicting ribosome stalling, 21st Annual International Conference on Research in Computational Molecular Biology, RECOMB 2017, Hong Kong, China, 2017.3.3-3.7

三、获得学术奖励（仅不列此项时可删除该标题，标题序号按实际情况编排）

1. Tao Jiang, Fellow of the Association for Computing Machinery (ACM), 2007 -.
2. Tao Jiang, Fellow of the American Association for the Advancement of Science (AAAS), 2006 -.
3. Tao Jiang, Japan Society for the Promotion of Science Research Fellowship, 1996.

除非特殊说明，请勿删除或改动简历模板中蓝色字体的标题及相应说明文字

参与者 简历

钟坚成，湖南师范大学，工程与设计学院，副教授

教育经历（从大学本科开始，按时间倒序排序；请列出攻读研究生学位阶段导师姓名）：

2011/09-2015/06，中南大学，信息科学与工程学院，博士，导师：潘毅

2004/09-2007/06，湖南师范大学，工程与设计学院，硕士，导师：杨家红

2000/09-2004/06，湖南师范大学，工程与设计学院，本科

科研与学术工作经历（按时间倒序排序；如为在站博士后研究人员或曾进入博士后流动站（或工作站）从事研究，请列出合作导师姓名）：

1. 2015/10-至今，湖南师范大学，工程与设计学院，副教授

2. 2009/10-2015/09，湖南师范大学，工程与设计学院，讲师

3. 2007/09-2009/09，湖南师范大学，工程与设计学院，助教

曾使用其他证件信息（申请人应使用唯一身份证件申请项目，曾经使用其他身份证件作为申请人或主要参与者获得过项目资助的，应当在此列明）

主持或参加科研项目（课题）及人才计划项目情况（按时间倒序排序）：

1. 国家自然科学基金青年项目，61502166，基于生物网络的共享肽归属及蛋白质定性算法研究，2016/01-2018/12，20万元，在研，主持

代表性研究成果和学术奖励情况（每项均按时间倒序排序）

（请注意：①投稿阶段的论文不要列出；②对期刊论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、期刊名称、发表年代、卷（期）及起止页码（摘要论文请加以说明）；③对会议论文：应按照论文发表时作者顺序列出全部作者姓名、论文题目、会议名称(或会议论文集名称及起止页码)、会议地址、会议时间；④应在论文作者姓名后注明第一/通讯作者情况：所有共同第一作者均加注上标“#”字样，通讯作者及共同通讯作者均加注上标“*”字样，唯一第一作者且非通讯作者无需加注；⑤所有代表性研究成果和学术奖励中本人姓名加粗显示。）

一、期刊论文（仅不列此项时可删除该标题）

1. 第一作者论文（仅不列此项时可删除该标题）

(1) **J. Zhong**, J. Wang*, X. Ding, Z. Zhang, M. Li, F. Wu, Y. Pan, Protein

inference from the integration of tandem MS data and interactome networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016, published online, DOI:10.1109/TCBB.2016.2601618

(2) **J. Zhong**, J. Wang*, W. Peng, Z. Zhang, M. Li, A feature selection method for prediction essential protein. Tsinghua Science and Technology, 2015, 20(5), 491-499

(3) **钟坚成**，彭玮*，海量种群基因表达式编程的内存删冗算法，计算机工程，2014, 40(9): 233-237

(4) **J. Zhong**, J. Wang*, W. Peng, Z. Zhang, Y. Pan, Prediction of essential proteins based on gene expression programming, BMC Genomics, 2013, 14(S4): S7

2. 既非第一作者又非通讯作者论文（仅不列此项时可删除该标题，序号按实际情况编排）

(1) Z. Zhang, J. Wang*, J. Luo, X. Ding, **J. Zhong**, J. Wang, F. Wu, Y. Pan, Sprites: detection of deletions from sequencing data by re-aligning split reads, Bioinformatics, 2016, 32 (12): 1788-1796

(2) J. Wang*, **J. Zhong**, G. Chen, M. Li, F. Wu, Y. Pan, Clusterviz: a cytoscape app for clustering analysis of biological network, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12(4), 815-822

(3) X. Tang, J. Wang*, **J. Zhong**, Y. Pan, Predicting essential proteins based on weighted degree centrality, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014, 11(2): 407-418

二、授权发明专利（仅不列此项时可删除该标题，标题序号按实际情况编排）

(1) 王建新，**钟坚成**，李敏，基于蛋白质相互作用网络和蛋白质组学的蛋白质鉴定方法，2017.02.15，中国，ZL201410399487.5

附件信息

序号	附件名称	备注	附件类型
1	H-PoP and H-PoPG	Minzhu Xie, et al. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics, 2016, 32(24):3735-3744.	代表性论著
2	LGH	Minzhu Xie, Jianxin Wang, and Xin Chen. LGH: A Fast and Accurate Algorithm for Single Individual Haplotyping Based on a Two-Locus Linkage Graph. IEEE/ACM Trans Comput Biol Bioinform, 2015, 12(6):1255-66.	代表性论著
3	BOP	Minzhu Xie, Jianxin Wang, and Tao Jiang. A fast and accurate algorithm for single individual haplotyping. BMC Systems Biology, 2012, 6(Suppl 2):S8.	代表性论著
4	EDCF	Minzhu Xie, Jing Li, and Tao Jiang. Detecting genome-wide epistases based on the clustering of relatively frequent items. Bioinformatics, 2012, 28(1):5-12.	代表性论著
5	ISMB2008-Xie	Minzhu Xie, Jing Li, and Tao Jiang. Detecting genome-wide epistases based on the clustering of relatively frequent items. Bioinformatics, 2012, 28(1):5-12.	代表性论著

签字和盖章页(此页自动生成,打印后签字盖章)

申 请 人： 谢民主 依托单位： 湖南师范大学

项目名称： 多倍体单体型从头组装算法研究

资助类别： 面上项目 亚类说明：

附注说明： 常规面上项目

申请人承诺：

我保证申请书内容的真实性。如果获得资助，我将履行项目负责人职责，严格遵守国家自然科学基金委员会的有关规定，切实保证研究工作时间，认真开展工作，按时报送有关材料。若填报失实和违反规定，本人将承担全部责任。

签字：

项目组主要成员承诺：

我保证有关申报内容的真实性。如果获得资助，我将严格遵守国家自然科学基金委员会的有关规定，切实保证研究工作时间，加强合作、信息资源共享，认真开展工作，及时向项目负责人报送有关材料。若个人信息失实、执行项目中违反规定，本人将承担相关责任。

编号	姓名	工作单位名称	证件号码	每年工作时间（月）	签字
1	姜涛	清华大学	HB675861	1	
2	钟坚成	湖南师范大学	430105198112060512	4	
3	周建宇	清华大学	130903199302150337	10	
4	熊袁鹏	清华大学	360425199407282516	10	
5	叶云洋	湖南师范大学	36232519821014295X	10	
6	彭哲也	湖南师范大学	430922199208240010	10	
7	周佩霞	湖南师范大学	430124199002094627	10	
8	唐紫珺	湖南师范大学	431103199305310023	10	
9	喻昕	湖南师范大学	430124199312084985	10	

依托单位及合作研究单位承诺：

已按填报说明对申请人的资格和申请书内容进行了审核。申请项目如获资助，我单位保证对研究计划实施所需要的人力、物力和工作时间等条件给予保障，严格遵守国家自然科学基金委员会有关规定，督促项目负责人和项目组成员以及本单位项目管理部门按照国家自然科学基金委员会的规定及时报送有关材料。

依托单位公章	合作研究单位公章1	合作研究单位公章2
日期：	日期：	日期：