****

密级：\_\_\_\_\_\_\_\_\_\_

本科生毕业设计(论文)

癌症组织监测与分析系统的设计与实现

题 目： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

黄凯辉

作 者： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

41467024

学 号： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

化学与生物工程学院

学 院： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

生物技术专业

专 业： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

成 绩： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2018年 06 月

**本科生毕业设计(论文)**

癌症组织检测与分析系统的设计与实现

**题 目:**

Design and Implementation of Cancer

**英文题目：**

Organization Monitoring and Analysis System

化学与生物工程学院

**学 院:**

生物技术1401

**班 级:**

黄凯辉

**学 生:**

41467024

**学 号:**

教师博士后

田澍

**指导教师:**  **职称：**

**指导教师:**  **职称：**

声 明

本人郑重声明：所呈交的论文是本人在指导教师的指导下进行的研究工作及取得研究结果。论文在引用他人已经发表或撰写的研究成果时，已经作了明确的标识；除此之外，论文中不包括其他人已经发表或撰写的研究成果，均为独立完成。其它同志对本文所做的任何贡献均已在论文中做了明确的说明并表达了谢意。

学生签名：\_\_\_\_\_\_\_\_\_\_\_ 年 月 日

导师签名：\_\_\_\_\_\_\_\_\_\_\_ 年 月 日

**毕 业 设 计（论 文）任 务 书**  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

一、学生姓名：                 学号：   
二、题目:   
三、题目来源：真实 、    自拟   
四、结业方式：设计 、    论文   
五、主要内容：  
  
  
  
六、主要（技术）要求：  
  
  
七、日程安排：  
  
  
  
  
八、主要参考文献和书目：

指导教师签字：             年   月   日

学 生 签 字：             年   月   日

系（所）负责人章：             年   月   日

摘 要

肿瘤纯度和倍性对肿瘤样本的下一代测序序列分析有重要作用，并可能改变其结果的生物学和临床解释。目前存在几种专门用于评估来自癌症基因组图谱肿瘤 - 正常细胞全基因组测序（WGS）数据的肿瘤纯度或倍性的计算方法，但仍然缺乏一种能够在广泛测序范围内，肿瘤纯度水平和肿瘤内异质性水平下，准确，快速且完全自动化的方法。

我们发明了一种称为Accurity的算法，通过体细胞拷贝数（SCNA）和杂合生殖系统单核苷酸变体（HGSNV）联合建模，推断肿瘤正常WGS数据的肿瘤纯度，肿瘤细胞倍性和绝对等位体拷贝数以用于体细胞拷贝数改变（SCNA）。 通过运算模拟测序数据和真实测序数据，结果表明，即使在低纯度，高倍性和低覆盖率环境中，Accurity仍具有高度准确性和稳健性。其中，几种现有方法相对表现不佳。 考虑到推断肿瘤的纯度和倍性时，Accurity显着增加了不同拷贝数之间的信号/噪声间隔。 我们希望Accurity能够用来进行癌症诊断性生物标志物的鉴定。

关键词： 二代测序，癌症，染色体倍性，肿瘤纯度，

Design and Implementation of Cancer Organization Monitoring and Analysis System

Abstract

**Motivation**

Tumor purity and ploidy have a substantial impact on next-gen sequence analyses of tumor samples and may alter the biological and clinical interpretation of results. Despite the existence of several computational methods that are dedicated to estimate tumor purity and/or ploidy from The Cancer Genome Atlas (TCGA) tumor-normal whole-genome-sequencing (WGS) data, an accurate, fast and fully-automated method that works in a wide range of sequencing coverage, level of tumor purity and level of intra-tumor heterogeneity, is still missing.

**Results**

We describe a computational method called Accurity that infers tumor purity, tumor cell ploidy and absolute allelic copy numbers for somatic copy number alterations (SCNAs) from tumor-normal WGS data by jointly modelling SCNAs and heterozygous germline single-nucleotide-variants (HGSNVs). Results from both in silico and real sequencing data demonstrated that Accurity is highly accurate and robust, even in low-purity, high-ploidy and low-coverage settings in which several existing methods perform poorly. Accounting for tumor purity and ploidy, Accurity significantly increased signal/noise gaps between different copy numbers. We are hopeful that Accurity is of clinical use for identifying cancer diagnostic biomarkers.

Key Words： next-gen sequence analyses， tumor， intra-tumor heterogeneity

目 录

[摘 要 I](#_Toc517166584)

[Abstract III](#_Toc517166585)

[注释说明清单 VII](#_Toc517166586)

[1 引 言 1](#_Toc517166587)

[2 文献综述 3](#_Toc517166588)

[2.1 癌症研究现状 3](#_Toc517166589)

[2.2 二代测序简介及存在的问题 5](#_Toc517166590)

[2.3 国内外对检测癌症组织纯度与倍性的研究现状 5](#_Toc517166591)

[3 系统开发工具 7](#_Toc517166592)

[3.1 Rust语言 7](#_Toc517166593)

[3.1.1 Rust语言主要特点 7](#_Toc517166594)

[3.1.2 使用到rust语言的一些基本库 8](#_Toc517166595)

[3.2 生物信息编码文件vcf, bam（sam） 8](#_Toc517166596)

[3.3 利用Pyflow流程控制 11](#_Toc517166597)

[4 算法 13](#_Toc517166598)

[4.1 功能模块总图 13](#_Toc517166599)

[4.2 算法内容 13](#_Toc517166600)

[5 实验 22](#_Toc517166601)

[参考文献 23](#_Toc517166602)

[在学取得成果 27](#_Toc517166603)

[致 谢 29](#_Toc517166604)

注释说明清单

**Whole Genome Sequencing(WGS)**： 使用二代测序技术的全基因组测序。

**read**： 高通量测序平台产生的测序序列。

**测序深度**：测序得到的碱基（bp）总量与基因组（Genome）大小的比值，它是评价测序量的指标之一。

**window（窗口）**：按照一定长度划分的基因组片段，该长度代表 window 大小。本方法中 window 大小可由使用者自由设置，通常设置为几百碱基。一个大基因组片段 S 可以包含大量 window。

**Tumor Read Enrichment（TRE）**： 癌症片段读长富集程度𝑒𝑠，指癌症样本中某一片段 S 内 read 数量与相应正常样本中对应片段 read 数量的比值，定义公式如下：

上式中，和分别表示在癌症样本中覆盖片段s的read数量和相匹配的正常样本中覆盖片段s的read数量，𝑁𝑡表示癌症样本全基因组测序获得 read总数量，𝑁𝑛表示相应正常样本全基因组测序获得 read 总数量。

**Heterozygous Germline Single Nucleotide Variants（HGSNV）**：杂合生殖系细胞单碱基变异，（在文件）由于人类染色体属于二倍体，体细胞均由胚胎细胞发育而来，而生殖细胞中HGSNV 位点只有两种碱基类型 A 和 B，其中一种来源于父本，另一种来源于母本。

**Major Allele Fraction(MAF)**：主要等位基因(allele)分数，本发明中使用的 HGSNV只有两种等位基因，一种等位基因与参考基因组相同，另一种与参考基因组不同。这两种等位基因型分数的计算方法为覆盖某一等位基因的 read 数量除以覆盖该位点总 read 数量的比值，MAF 就是两种等位基因分数中的较大值。计算公式如下所示，为包含与参考基因组相同等位基因的 read 数量，为包含另一种等位基因的 read 的数量，表示覆盖该 HGSNV 位点的总 read 数量， 𝐶为该 HGSNV 的 MAF 值。MAF 是相对于 HGSNV 的概念，本发明中“片段的 MAF”指片段内所有 HGSNV 的 MAF 均值，“peak 的 MAF”指peak 中所有片段包含的 HGSNV 的 MAF 均值。

**Major allele copy number（Maf）**： 主要等位基因拷贝数，指在拷贝数为 i 的片段中，主要等位基因拷贝数的取值，它的取值范围为大于等于𝑖的整数。

**Peak**：指基因组所有 window 的 TRE 分布中，聚集在一起的 TRE 簇。如附图 1 所示，图 A 表示基因组上所有 window 的 TRE 分布，纵轴表示对应某 TRE 位点的 window总数量，该图为基因组 GC 含量校正之前的 TRE 分布，图 B 表示 GC 含量校正之后的 TRE分布，图 B 中可以看到 window 明显以簇聚集，本方法将通过类自回归模型鉴定出来的TRE 簇定义为 peak，本质上是具有相同拷贝的基因组片段内 window 的聚集癌症样本： 指从某患癌症的个体身上取下的癌症组织，它包含了一部分癌症细胞和一部分正常细胞。

**P**： 指两个相邻 peak 之间的间距，由于 peak 是一个簇，这里 peak 的 TRE 由 peak 的 TRE 均值来表示，所以实际上是两个相邻 peak 的 TRE 均值的差。由于 peak 是具有相同拷贝数基因组片段内 window 的聚集，所以这里也表述为相邻拷贝数片段 TRE 的差值。

1. 引 言

癌症是一组异质疾病，每种疾病都有自己的生物特征。发现这些生物特征可能产生癌症治疗剂的高度信息标记和靶标[1,2,3,4]。最近，新一代测序技术（NGS）使科学家能够在全基因组范围内搜索这些癌症特征[5,6,7]。然而，肿瘤纯度（以异种肿瘤样本中癌细胞的比例来衡量）和肿瘤细胞倍性（癌症基因组的平均拷贝数）对肿瘤样本的NGS分析具有实质性影响，并可能改变生物学和临床解释的结果[8,9,10]。在本文中，我们区分肿瘤细胞倍性（短为肿瘤倍性）和肿瘤样品倍性（肿瘤样品中正常和癌细胞之间的平均值，如方法中所述）。传统上，病理学家的任务是通过视觉检查肿瘤样品来评估肿瘤纯度和倍性。基因组技术在过去十年的发展为从计算机推断肿瘤纯度和来自基因组数据的倍性打开了大门。最近，肿瘤正常对测序在研究人员对癌症基因组进行分析时获得了显着的推动力，因为其在仅有肿瘤的测序中提高了统计功效[11,12]。癌症基因组图谱（TCGA）包含近千个由高覆盖率（> 30x）全基因组测序（WGS）和另外一千个由低覆盖率（6-8×）WGS 。大量的计算方法[13,14,15,16,17,18,19]已被开发用于推断肿瘤正常对WGS数据的肿瘤纯度和倍性。

估计肿瘤纯度和倍性依赖于可以区分肿瘤细胞与肿瘤样品中正常细胞的统计信号。肿瘤NGS数据中的统计分化主要来自两种类型的遗传变异。一种类型的事件是体细胞拷贝数改变（SCNA）。比较肿瘤样品的SCNA位点与其匹配的正常样品的测序范围构成了统计学差异。第二种是单核苷酸变异体（SNVs）。比较肿瘤样品的SNV基因座的等位基因测序覆盖率与其匹配的正常样品的等位基因测序覆盖率构成第二统计学差异。根据这两类事件的覆盖率信息如何用于估计肿瘤纯度和倍性，现有的计算方法可以大致分为三类。第一类仅利用SCNA的覆盖范围信息[20,21]。第二类仅利用SNV的覆盖信息[22,23]。第三类利用这两个信息[24]。

影响第一类和第二类方法的一个问题是可识别性问题，其中肿瘤纯度和肿瘤细胞倍性的不同组合可以同样很好地解释观察到的数据[25]。对于仅利用SCNA的覆盖率信息的方法，（30％，3），肿瘤纯度= 30％和肿瘤细胞倍性= 3的组合可以解释这种肿瘤样品的测序覆盖率，因为根据公式(1)，(在第四章：算法)这些组合导致相同的肿瘤样本倍性= 2.3。类似地，仅利用SNV的覆盖率信息的方法遭受同样的问题。为了规避可识别性问题，这些方法做出了明确或隐含的假设，有助于将候选结果缩小到一个解决方案。例如，使用体细胞突变（一种SNVs）的B等位基因频率（BAF）来估计肿瘤纯度的方法PurityEst[26]）有效地假定肿瘤细胞倍性等于2.CNAnorm[27]，这是最接近二倍体的解决方案。 ABSOLUTE[28]除SCNA的覆盖率信息外，还包含核型数据。 Oesper等人提出[29]将所有最优解或极限输出到克隆肿瘤群体基线拷贝数的解决方案。

正如在基础代数中需要两个方程来解决双变量系统一样，通过结合SCNA和SNV的覆盖信息，第三类方法可以从根本上解决这个可识别性问题[30]。考虑到这些方法中的一些，即MixClone[31]，甚至考虑肿瘤内异质性（IRH）[32]。正如所证明的关于肿瘤纯度和倍性的知识可能对SCNA的检测具有显着影响，这对于癌症进展是重要的[33]。检测SCNA的能力高度依赖于肿瘤的纯度。在低纯度肿瘤样品中，来自非癌细胞的大部分拷贝中性DNA显着降低了SCNA的信噪比。假设肿瘤纯度为100％，肿瘤纯度不可知的SCNA调用者几乎肯定会缺乏动力[33,34]。肿瘤倍性的知识对于检测SCNA也是至关重要的。一些非整倍体癌症样本的倍性可能远不止两种。使用全基因组平均拷贝数作为基线的方法可以将拷贝中性区域视为正常的缺失和扩增。相反，当肿瘤倍性较低时，这些方法可以将拷贝中性区域称为正常的扩增和中间缺失。

我们将Accurity应用于模拟数据，并证明Accurity可以在各种设置下产生准确和稳健的估计值：重点在于SCNA或染色体SCNA或全基因组重复（WGD），低覆盖率或高覆盖率。我们还将Accurity应用于来自数十个TCGA样品的真实测序数据。 Accurity的纯度估计与组织学估计高度一致。考虑到肿瘤的纯度和倍性，Accurity显着增加了不同拷贝数之间的信号/噪音差距，并且可以帮助识别复杂的SCNAs，这对癌症诊断具有极大推动，启发作用

1. 文献综述
   1. 癌症研究现状

癌症是严重威胁人类生命和社会发展的重大疾病，近几十年来，随着疾病模式的转变和人口老龄化趋势，我国癌症负担日益增加，癌症防治面临严峻的形势。如何科学地预防，控制癌症，已成为全球最重要的公共卫生问题之一。

自上世纪起，癌症发病人数日益增加，政府也对癌症日益重视。我国三次全国范围内的死因调查数据显示，近30 年，中国癌症在死因中的构成比由20 世纪70 年代的10.13% 上升至22.32%，死亡率从73.99/10万上升至135.88/10 万。在城市地区，癌症列居全死因的第一位，而在农村地区，列居全死因的第二位。癌症死亡率升高的主要原因与人口老龄化、微生物感染、吸烟、饮食变化、活动减少及肥胖增加等相关。根据“1992 年第三次全国营养调查” 及“2002 年中国居民营养与健康调查” 资料结果，研究发现十年间我国居民的超重率和肥胖率分别上升了38.6% 和80.6%。

就全球而言，全球癌症总死亡数2013 年比1990年增加45. 47%，增加256 万人; 2013 年与1990 年相比，年龄标化死亡率减少14. 97%。1990 年至2013 年中国的癌症总死亡数增加率稍低于全球水平，年龄标化死亡率减少比例大于全球水平。

2013 年排名前9 位癌症按全球的死亡数和发病数由大到小排序，如表2 和表3 所示，中国所有癌症( 除外非黑色素瘤皮肤癌) 、肺癌、胃癌、肝癌和食管癌的年龄标化死亡率均明显高于全球水平; 肺癌、胃癌和肝癌的年龄标化发病率均明显高于全球水平。

总体而言，中国癌症新发病例增加幅度高于全球平均水平，主要是由于期望寿命增加较快、人口老龄化迅速所致，消除人口老龄化的影响后，中国癌症发病率目前仍低于全球平均水平; 而中国癌症的死亡人数比例远高出中国癌症发病人数占全球的比例，主要是癌症患者发现病情时以中晚期居多，治疗效果欠佳。

癌症的发病机理一般是源自遗传的基因突变或获得性的体细胞突变。因此，基因组学是研究癌症必不可少的一个方面。2008 年，利用高通量测序技术第一个急性髓性白血病患者的基因组被测定。截至目前，国际癌症基因组联盟( international cancer genomeconsortium， ICGC) 已经向科研界公布了超过1 万个癌症基因组的数据。而20 年前的人类基因组计划，需要国际合作花费10 年时间，耗资38 亿美元才能完成一个人类基因组的测序［6］。随着测序技术的进步和成本的降低，更多的癌症基因组将被测序完成。这些海量数据的获得，不但大大丰富了癌症基础生物学研究的数据和知识，并且大大增加了在其中寻找有关癌症预防，诊断，预后和治疗信息的机会。

到目前为止，大多癌症基因组测序相关的研究都是为了实现以下四个特定目标或其中部分目标: 发现驱动突变、识别体细胞突变的特征、表征克隆进化特点、推进个性化医疗。第一个方面，确定哪些突变有可能导致癌症表型变化是癌症基因组测序研究最普遍的目的。发现驱动突变可以提高我们对癌症基础生物学的理解，并由此推动新的治疗方法的发现和发展。以zeste2 基因( EZH2) 为例，二代测序发现在淋巴瘤中，EZH2 基因在临床上具有非常显著的突变率，从而促使EZH2 基因的功能得以解析，并使其成为一个潜在的治疗靶点; 第二个方面，发现体细胞突变的特征也有助于获得更多关于癌症基础研究的知识。研究者可以揭示整个突变和DNA 修复机制的情况和特征。其中，kataegis和chromothripsis这两种突变类型就是在大量的癌症基因组测序中发现的。第三个方面，表征克隆进化的特点是一个非常重要的方面，尤其是对于癌症治疗，同时通过测序可以在核酸水平揭示克隆进化的特点。例如，多数细胞可能具备相同的固有的药物耐受突变类型，然而其中一小部分细胞或产生了新的突变的细胞会产生获得性的药物耐受情况。如有文章报道，大鼠肉瘤中v-Ki-Ｒas2 基因的突变导致大鼠对EGFＲ 靶向治疗的耐受。第四个方面，即推进个性化医疗，是目前癌症基因组测序比较清晰的应用方向。个体化用药的目的就是为了有目的的对患者选择合适的药物，合适的剂量和治疗时间，从而降低药物毒副作用。成神经管细胞瘤是深入研究个性化用药的模式肿瘤类型，因为它是一个异源性的肿瘤类型，在整个的生存率和分子特征上来说，个体差别都比较大。而且，积极的治疗方案虽然会降低死亡率，但会极大地升高发病率。因此，找到合适的适用于积极治疗方案的患者对于提高存活的成神经管细胞瘤患者的生活质量具有重要的作用。除了以上提到的四点，还有很多癌症基因组测序可以解决的问题。比如，与儿童癌症相关的新的胚系变异，可以通过对先证者的后代和其父母的研究发现，并且这些特定目标之间是相通的，例如研究克隆进化同样有助于驱动突变的发现。

* 1. 二代测序简介及存在的问题

二代测序技术(Next-generation sequencing technology)又叫做大规模平行测序技术(massively parallel signature sequencing,MPSS),是相对于一代测序技术(Sanger sequencing)而言近年来新发展起来的测序技术。根据测序内容的不同，二代测序技术可以大致分为以下几个内容：

（一）全基因组测序(Whole genome sequencing, WGS)

利用整个基因组的ＤＮＡ建库，测得整个基因组的全部碱基序列。全基因组测序的覆盖面广，有助于对整个基因组内的情况的全面了解，但是测序数据大，费用较高，对算法和数据分析运行环境的要求较高，并且目前对基因组中大部分区域了解较少，不能有效利用分析结果。

（二）全外显子组测序(Who exome sequencing, WES)

在建库时利用特殊方法提取外显子部分并对其进行测序。外显子仅占基因组的２％，得到的测序数据量小，并且外显子组与细胞中执行功能的蛋白质联系紧密，更有利于分析与蛋白编码相关的事件，另一方面也可能因为覆盖范围小影响分析。

（三）转录组测序(Transciptome sequencing, RNA-seq)

建库时仅选择以poly(A)结尾的RNA序列进行测序，这部分RNA大多数为mRNA，还包括部分不成熟的RNA和LnkRNA。mRNA的组成受到细胞类型和时期的影响，细胞特异性较强，细胞之间差异较大。但是mRNA与蛋白质直接相关，且覆盖范围较全基因组测序和外显子测序更小，容易达到满意的测序深度，因此被广泛应用于各类分析中。

* 1. 国内外对检测癌症组织纯度与倍性的研究现状

癌症的研究是生命医学中的重要研究领域，并对人类健康生活有重大影响。癌症是一类细胞恶性增殖的疾病，因其病理十分复杂，人类尚无法攻克这类疾病。二代测序（next generation sequencing）为快速检测病人遗传信息提供了可能。然而测序需要从病人组织中提取样本，但通常癌症组织并不是只单纯地包含癌症细胞，它还有非常丰富的微环境。癌症细胞微环境指包围或伴随癌症细胞的正常细胞(noncancerous cells)环境。癌细胞样本提取时，这些微环境会和癌细胞一起被提取，并会伴随癌细胞一起被测序[35]。癌症细胞在癌症样本中的比例被定义为癌症样本的纯度。癌症基因组通常包含着大量体细胞序列拷贝数变异，这些变异主要由基因组片段扩增或删除造成。识别特定肿瘤基因组的基因组片段拷贝数变化，是癌症基因组研究的一个重要课题。要准确鉴定基因组片段拷贝数具有一定挑战，因为癌症片段拷贝数主要由两个因素混合决定，一是癌症样本纯度，即癌症细胞在癌症样本中所占比例，二是染色体倍性[36,37]。传统中鉴定癌症样本纯度和染色体倍性的方法是使用实验技术，如定量图像分析[38,39]或单细胞测序[40]。但是在大型项目中，这样的方法会耗费大量人力、资金和时间。随着测序技术地发展，测序数据地快速增长，以及测序数据分析技术的积累，各种各样的癌症样本纯度算法被提出，并开发出了相对应的软件。

基于基因组片段拷贝数变异和等位频率（突点的 基于因组片段拷贝数变异和等位频率（突点的B-等位基因（ B-allele）） 的计算方法被相继提出。基于等位因频率有 PurityEst[41]和 PurBayes[42]，主要是依赖于随着肿瘤样本纯度和肿瘤基因组倍性的不同，等位基因的频率会有所不同。基于拷贝数变异的方法有CNAnorm[43]、THet A[44]、和 ABSOLUTE[45]等。然而这两种方法都有不同程度的问题，使用等位基因频率的方法由于数据量的问题会有较大的误差，而运用拷贝数变异的方法虽然较稳定，但却无法区分样本纯度和染色体倍性的补偿效应，即存在识别问题。以上基于片段拷贝数的软件都没有解决这一问题，CNAnorm 倾向于选择染色体倍性离二倍体最近的解决方案，ABSOLUTE 结合了其他的经验数据，THet A 则直接将所有可能结果都列出来了。更优的方案应该是结合等位基因频率信息和片段拷贝数信息共同计算肿瘤样本纯度。Py LOH[46]，patchwork[47]使用了基因组上杂合 SNV（单核苷酸变异）位点的频率信息，和基因组片段的拷贝数。Py LOH 一定程度上解决了“识别困难”的问题，可以更合理给出唯一解决方案。但是其准确性较差，特别是遇到基因组中存在亚克隆（subclone）的情况下。Patchwork 同时使用了两种信息，但是在计算基因型的中间步骤中，需要人工识别，人工判断的结果缺乏准确性，并且这种半自动化软件给应用带来很多不便。

如何充分利用现有的二代测序数据准确计算癌症样本纯度和癌症细胞基因组倍性问题仍然是一项具有挑战性的工作。

1. 系统开发工具
   1. Rust语言

Rust 是 Mozilla 开发的注重安全、性能和并发性的编程语言。R是针对多核体系提出的语言，并且吸收一些其本程序选择他动态语言的重要特性，比如不需要管理内存，比如不会出现 Null 指针等。

本程序选择使用rust语言作为主要编写语言，并辅以python,C++语言共同编写。

* + 1. Rust语言主要特点

Rust语言是一个极其讲究精致的系统级别语言，由web语言的领军人物Brendan Eich（js之父），Dave Herman以及Mozilla公司的Graydon Hoare 合力开发。创建这个新语言的目的是为了解决一个顽疾：软件的演进速度大大低于硬件的演进，软件在语言级别上无法真正利用多核计算带来的性能提升。Rust是针对多核体系提出的语言，并且吸收一些其他动态语言的重要特性，比如不需要管理内存，比如不会出现Null指针等等，其主要有以下优点：

1. Rust语言无需竞争的并发，Rust能够确保在并发编程中的数据安全，某个时间点同时只能有多个读操作或一个写操作被允许访问共享数据。

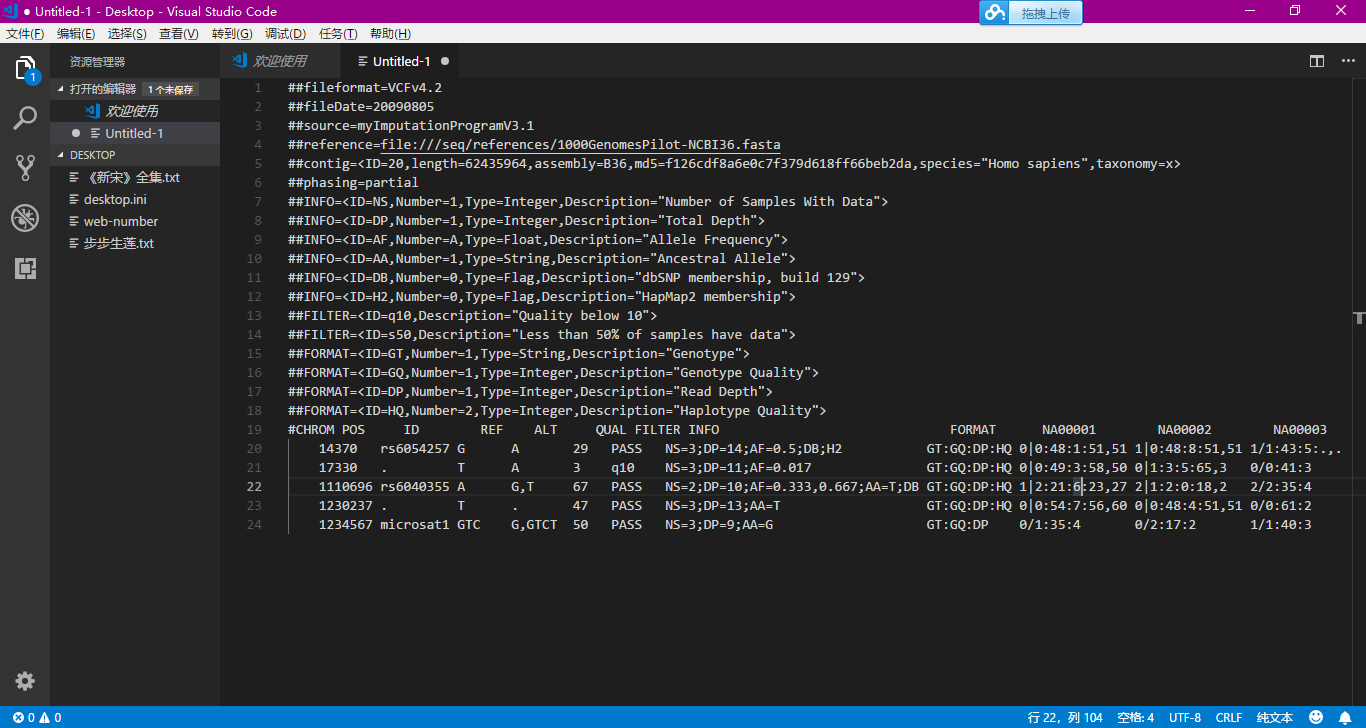
2. Rust语言提供代数数据类型，除了tuple struct类型以外，Rust还提供enum和模式匹配，这些都是高级类型系统才拥有的功能如今出现在系统级的编程语言中。

3. Rust语言是基于继承的组合，Rust能基于继承实现组合，Trait是Rust泛型中重要角色。

在我们的软件开发过程中，需要多次对癌症样本全基因组数据进行数据处理，其中样本数据一般在5G~30G之间（例如TCGA-CM-4746-01A肿瘤样本数据大小为12.5G，该样本数据以100bp~300bp基因片段read为一行记录了总大小为30亿bp的癌症样本的所有序列和序列信息。根据测序时对read分割不同而使总文件大小不同）。由于所处理的数据内容较大，内容繁琐，需要一种高性能，高效率的程序语言进行高性能计算。为此，我们把目光瞄向了rust语言和Go语言。

* + 1. 使用到rust语言的一些基本库
  1. 生物信息编码文件vcf, bam（sam）

VCF是用于描述SNP（单个碱基上的变异），INDEL（插入缺失标记）和SV（结构变异位点）结果的文本文件。在GATK软件中得到最好的支持，VCF文件分为两部分内容：以“#”开头的注释部分，通过注释部分的阅读可以更好地了解文件各行各列的内容；没有“#”开头的主体部分。 主体部分中每一行代表一个Variant的信息。



现对一些重要的列解释如下

**CHROM**：表示变异位点是在哪个contig 里call出来的，如果是人类全基因组的话那就是chr1…chr22，chrX,Y,M。

**POS**： 变异位点相对于参考基因组所在的位置，如果是indel，就是第一个碱基所在的位置。

**ID**： variant的ID。 如果call出来的SNP存在于dbSNP数据库里，就会显示相应的dbSNP里的rs编号；若没有，则用’.’表示其为一个novel variant。

**REF**和**ALT**： 在这个变异位点处，参考基因组中所对应的碱基和研究对象基因组（Variant）中所对应的碱基。

**QUAL**： Phred格式(Phred\_scaled)的质量值，可以理解为所call出来的变异位点的质量值。表 示在该位点存在variant的可能性；该值越高，则variant的可能性越大；

计算方法：

① Q=-10\*lgP，Q表示质量值；P表示这个位点发生错误的概率。

②Phred值Q = -10 \* lg (1-p) ，p为variant存在的概率;

通过计算公式可以看出值为10的表示错误概率为0.1，该位点为variant的概率为90%。

同理，当Q=20时，错误率就控制在了0.01。

**FILTER**： 使用上一个QUAL值来进行过滤的话，是不够的。理想情况下，QUAL这个值应该是用所有的错误模型算出来的，这个值就可以代表正确的变异位点了，但是事实是做不到的。因此，还需要对原始变异位点做进一步的过滤。无论你用什么方法对变异位点进行过滤，过滤完了之后，在FILTER一栏都会留下过滤记录，如果是通过了过滤标准，那么这些通过标准的好的变异位点的FILTER一栏就会注释一个PASS，如果没有通过过滤，就会在FILTER这一栏提示除了PASS的其他信息。如果这一栏是一个“.”的话，就说明没有进行过任何过滤。

**INFO**： 这一行是variant的详细信息。

SAM是一种序列比对格式标准， 由sanger制定，是以TAB为分割符的文本格式。主要应用于测序序列mapping到基因组上的结果表示，当然也可以表示任意的多重比对结果。当测序得到的fastq文件map到基因组之后，我们通常会得到一个sam或者bam为扩展名的文件。SAM的全称是sequence alignment/map format。而BAM就是SAM的二进制文件(B取自binary)。

SAM由头文件和map结果组成。头文件由一行行以@起始的注释构成；而map结果是类似下面的东西：在SAM输出的结果中每一行都包括十二项通过Tab分隔，从左到右分别是

1：序列的名字(ID)

2：概括出一个合适的标记，各个数字分别代表

|  |  |
| --- | --- |
| 1 | 序列是一对序列中的一个 |
| 2 | 比对结果是一个pair-end比对的末端 |
| 4 | 没有找到位点 |
| 8 | 这个序列是pair中的一个但是没有找到位点 |
| 16 | 在这个比对上的位点，序列与参考序列反向互 |
| 32 | 这个序列在pair-end中的的mate序列与参考序列反响互补 |
| 64 | 序列是 mate 1 |
| 128 | 序列是 mate 2 |

假如说标记为以上列举出的数目，就可以直接推断出匹配的情况。假如说标记不是以上列举出的数字，比如说83=（64+16+2+1），就是这几种情况值和。

3：参考序列的名字

4：在参考序列上的位置(头or尾？)

5：mapping qulity 越高则位点越独特。bowtie2有时并不能完全确定一个短的序列来自与参考序列的那个位置，特别是对于那些比较简单的序列。但是bowtie2会给出一个值来显示出 这个段序列来自某个位点的概率值，这个值就是mapping qulity。

Mapping qulity的计算方法是：Q=-10log10p，Q是一个非负值，p是这个序列不来自这个位点的估计值。

假如说一条序列在某个参考序列上找到了两个位点，但是其中一个位点的Q明显大于另一个位点的Q值，这条序列来源于前一个位点的可能性就比较大。Q值的差距越大，这独特性越高。

Q值的计算方法来自与SAM标准格式，请查看SAM总结。

6：代表比对结果的CIGAR字符串，如37M1D2M1I，这段字符的意思是37个匹配，1个参考序列上的删除，2个匹配，1个参考序列上的插入。M代表的是alignment match(可以是错配)

7：mate 序列所在参考序列的名称

8：mate 序列在参考序列上的位置

9：计出的片段的长度，当mate 序列位于本序列上游时该值为负值。

10：read的序列

11：ASCII码格式的序列质量

12：可选的区域

|  |  |
| --- | --- |
| AS:i | 匹配的得分 |
| XS:i | 第二好的匹配的得分 |
| YS:i mate | 序列匹配的得分 |
| XN:i | 在参考序列上模糊碱基的个数 |
| XM:i | 错配的个数 |
| XO:i | Gap open的个数 |
| XG:i | gap 延伸的个数 |
| NM:i | 编辑距离。但是不包含头尾被剪切的序列。一般来说等于序列中error base的个数 |
| YF:i | 说明为什么这个序列被过滤的字符串 |
| MD:Z | 代表序列和参考序列错配的字符串 |

* 1. 利用Pyflow流程控制

pyFlow是一个轻量级并行任务引擎，它是一个基于Flow概念设计的python模块，是在任务依赖关系图的上下文中管理任务的工具。它支持使用普通过程式Python代码定义分布式异步工作流应用程序。Pyflow深受AWS [Flow Framework for Java] [Java Flow]和[Flow Framework for Ruby] [Ruby Flow]的启发，但并未尝试与这两种框架兼容。

Flow的基本概念很简单，就是一个有向无环图（DAG），数据在节点间流动。

001306_IHSy_1450051

节点 Node

节点是组成流的主要单元，负责对流入节点的数据进行处理，并输出到后续节点进行进一步的处理。

端口 Port

每个节点拥有输入和输出端口，输入端口负责数据流入节点，输出端口负责数据流出节点。每个节点都可能拥有一个或者多个输入和输出端口。

连接 Link

一个节点的输出端口连接到另一个节点的输入端口，节点处理好的数据通过连接流入其后的节点。

Flow的实现

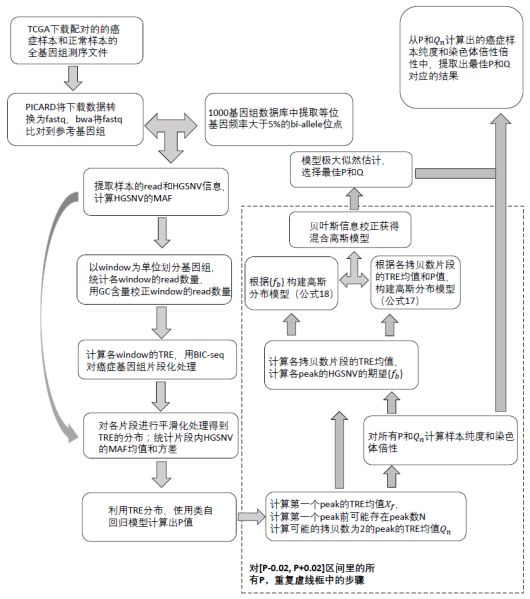
Pyflow对Flow的实现基本思路就是用一个Python的函数function实现一个节点，输入端口映射为函数的输入参数。输出端口映射为函数的返回值。

流中有一个节点被设置为终点节点（End Node），通过节点间的连接关系，以终点节点开始通过连接搜索所有的依赖关系（树形查找），得到一个节点运行的栈。005615_yCja_1450051

例如上图，我们就可以得到一个 [node1，node2, node3] 这样的栈。按顺序出栈的方式执行每一个节点的功能就可以运行整个流。（注意，这是一个简单版本的Flow的实现，仍然是一个批处理，不是streaming）

需要假定每一个节点的功能是无状态的，这样就可以利用输入输出端口对计算结果进行缓存，但输入值是已经运算过的值的时候，不需要运算，直接返回已经计算过的值。

1. 算法
   1. 功能模块总图



* 1. 算法内容

步骤 A：

获取配对的癌症组织样本和正常组织样本的全基因测序数据，并将测序数据比对到参考基因组；

步骤 B：

从步骤 A得到的比对结果文件中，提取 read位置和长度信息， HGSNV位点和覆盖 该位点的 read数量信息，计算所有HGSNV的 MAF，其中，计算公式如（1）所示：

 （1）

公式（1.1）中，𝑛𝑟为包含与参考基因组相同等位的 read数量， 𝑛𝑎为包含另一种等位基因的read的数量，𝑛𝑡表示覆盖该 HGSNV位点的总 read数量，𝐶为该HGSNV的MAF值；

步骤 C：

根据步骤 B得到的 read位置和长度信息，以window为单位统计各 window内包含的read 数量，使用基因组GC含量校正所有 window内 read 数量；

步骤 D：

使用步骤 C校正后的 read数量，使用公式（2）计算每一个window的 TRE，然后运用 TRE，通过 BIC-seq软件对基因组进行片段化，获得以拷贝数划分的基因组片段：

（2）

公式（2）中，和分别表示在癌症样本中覆盖片段s的read数量和在正常样本中覆盖片段s的read数量，表示癌症样本总read数量，表示相应正常样本总read数量，为TRE值；

步骤E：

以步骤D中BIC-seq处理后的基因组片段为单位，统计片段内所有window的TRE的均值、方差和该片段内window数量，根据均值和方差对基因组每个片段的window数量进行平滑化处理，使TRE的分布更均匀，然后将平滑化处理后所有片段的window分布汇总，得到基因组上window随TRE变化的分布结果；同时以片段为单位，计算片段中所有HGSNV的MAF的均值和方差；

步骤F：

使用如公式（3）、（4）所示的类自回归模型，计算相邻拷贝数片段内TRE的差值即𝑃，其中，遍历一定范围的P，计算Y(𝑃)，在Y(𝑃)的分布中，选择第二高峰内Y(𝑃)的最大值对应的𝑃作为P的计算结果：

(3)

(4)

公式（3）和（4）中，表示 0到𝑀𝑡之间的 TRE值；

t表示扩大了1000倍的TRE值；

𝑀𝑡表示TRE的最大值；

变量𝑃表示两个TRE位点的间隔；

𝐶()表示在TRE为的位点，对应的window数量；

C(𝑋𝑡 +1000×𝑃)表示在TRE为的位点对应的window数量；

𝑌(𝑃)表示在变量 𝑃下，类自回归模型的函数值；

步骤G：

根据步骤F得到的𝑃，计算TRE分布中第一个实际观测peak的TRE均值，然后计算在第一个实际peak之前最多可能存在理论之前最多可能存在理论peak的数量N，最后当第一个实际peak之前存在n个理论peak时，计算Q的值，以𝑄𝑛表示，其中步骤G包括：

G1：

根据步骤 F计算的 𝑃，使用公式（5），选取使公式（5）取最大值的𝑋𝑓作为第一个实际观测peak的TRE均值：

,0<<𝑀𝑡, 0<<𝑀𝑡 (5)

公式(5)中，i表示第i个peak;

表示在TRE为的位点对应的window数量;

n表示𝑀𝑡以内peak的最大数量;

𝑀𝑡表示TRE的最大值；

G2：

使用公式（6），根据步骤F计算的𝑃和步骤G1计算的，计算在之前最多可能存在的peak数量𝑁：

𝑁=𝑓𝑙𝑜𝑜𝑟() （6）

公式（6）中，表示第一个peak的均值，𝑃表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数；

G3：

利用步骤G2计算的N值，当n取0到N之间的整数时，使用公式（7）计算𝑄𝑛的值：

𝑄𝑛=−𝑛×𝑃+2×𝑃=+(2−𝑛)×𝑃，𝑛∈[0，𝑁] （7）

公式（7）中，n表示之前peak的数量，取值范围是0到N之间的整数，

P表示相邻拷贝数片段对应的peak之间的间距，

表示第一个实际观测peak的TRE均值，

𝑄𝑛表示在之前理论上存在n个peak时的Q值；

步骤H：

使用步骤F计算的P与步骤G计算的𝑄𝑛，使用公式（8）、（9）计算癌症样本纯度𝛾和染色体倍性𝜅：

𝛾= (8)

𝜅=2+ （9）

公式（8）、（9）中，𝛾表示样本纯度，𝜅表示染色体倍性， 由此对（𝑃，𝑄𝑛）得到对应的（𝛾，𝜅）；

步骤 I：

当n取[0,N]之间的某个整数值时，使用公式（10）计算第i个peak的 TRE均值：

𝑇𝑖=− 𝑛 ×𝑃+ 𝑖×𝑃= +(𝑖 −𝑛)×𝑃 𝑛 ∈[0，𝑁] （10）

公式（10）中，

n表示之前 peak的数量，取值范围是0到 N之间的整数， P表示 相邻拷贝数片段对应的 peak之间的间距，表示第一个实际观测 peak的 TRE均值， 𝑇𝑖表示第 i个 peak的 TRE均值 ，

对于落在 𝑇𝑖附近的片段，认为该片段具有拷贝数i；对于没有落在𝑇𝑖附近的片段，将其归类为亚克隆片段，在后续分析中剔除所有亚克隆片段；然后根据步骤 H计算的癌症样本纯度 𝛾和 peak对应的拷贝数，计算peak的 MAF的期望 𝑓𝑏，不同 peak的 MAF期望不同， 对基因组上的所有 peak，最终得到 MAF期望的集合{𝑓𝑏}；同时计算各个 peak的 TRE均值和方差或标准差；

步骤 J：

根据步骤 F计算的 P和步骤I计算的 {𝑓𝑏}构建如公式 （13）所示的用 “贝叶斯信息准则 ”校正后的混合高斯分布模型，然对极大似然估计；其中步骤 J包括如下几步：

J1：

以步骤 F计算的 P构建如公式（11）所示的高斯分布模型：

𝐿(𝑒𝑠; 𝛾,𝜅)= Π[Σ𝑝𝑖 × 1√2𝜋 𝜎𝑖 𝑒𝑥𝑝(− (𝑒𝑠− 𝑆𝑖)22𝜎𝑖2)𝐼𝑖=0] (17)𝑁𝑠 = 1 25

(11)

公式（11）中， 表示基因组片段 TRE的似然函数，

N表示基因组上的所有window的数量，

I表示基因组中所有片段的最大拷贝数，

表示拷贝数为 i的所有片段的TRE的标准差由步骤 I得到，

为第s个window的TRE观测值，

表示第 i个peak的TRE均值即步骤 I中的 𝑇𝑖，

表示第 s个 window的拷贝数为 i的权重，对所有的均取值为 1；

J2：

以步骤 I计算的𝑓𝑏构建如公式（12）所示的高斯分布模型 ：

(12)

公式（12）中， 𝐿(𝑓𝑠; 𝛾,𝜅)表示 HGSNV的似然函数，

M表示基因组中所有 HGSNV数量，

S表示第 S个 HGSNV，

I表示基因组中所有片段的最大拷贝数 表示基因组中所有片段的最大拷贝数 ;

表示拷贝数 5 为 i，主要等位基因的拷贝数为j的片段内 HGSNV的 MAF期望值，由步骤I得到；

𝑓𝑠表示该片段内所有 HGSNV的 MAF的观测值均，由步骤 E得到 ；

表示该片段内所有 HGSNV的 MAF观测值的标准差，由步骤 E得到 ；

表示在主要等位基因的拷贝数为 j时，高斯分布的权重对所有i和 j，取值均为 1，

𝑝𝑖表示第 S个 HGSNV所在片段的拷贝数为 i的权重，对所有i，𝑝𝑖取值均为1；

J3：

将（11）与（12）相加得到混合高斯模型，然后对混合模型进行BIC（Bayesian Information Criterion）校正得到最终混合模型如公式（ 13）： 𝐵𝐼𝐶(,𝑓𝑠; 𝛾,𝜅 )

=−2×𝑙𝑜𝑔𝐿(𝑓𝑠; 𝛾,𝜅)−2×𝑙𝑜𝑔𝐿(; 𝛾,𝜅)+𝐼 × 𝑙𝑜𝑔(𝑁) + 𝐽 × 𝑙𝑜𝑔(𝑀) (13)

公式（13）中,𝐵𝐼𝐶(,𝑓𝑠; 𝛾,𝜅 )表示混合模型的似然函数，I表示基因组中所有片段的最大的拷贝数,J是公式（18）中j的取值个数, N是基因组中 window的数量， M是基因组中 HGSNV的个数 ，

对[0, N]范围内的每一个整数值n，通过步骤 G得到 𝑄𝑛，或者 通过步骤 I得到所有 peak的 MAF期望的集合 {𝑓𝑏}，由一对（ P，{𝑓𝑏}）构建一个公式（ 19）所示的模型；

步骤 K：

以 0.001为分辨率，对 [P-m, P +m]区间的所有 P值，重复步骤 G~J，得到一系列不同的（ P，𝑄𝑛）与对应的 似然函数值，取最大 (P，𝑄𝑛)作为最合适的P和 Q值， m是 0到 0.5之间的一个值；

步骤 L：

查询步骤 H的结果， 找到在步骤 K得到的（ P，Q）下，对应的癌症样本纯度和染色。

所述步骤 A中， 采用 1000基因组计划第三期项目使用的参考基因组 hs37d5(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2\_reference\_assembly\_sequ25 ence/hs37d5.fa.gz)作为所述参考基因组，和 /或，比对软件使用 Burrows-Wheeler Aligner (BWA)，BWA)，比对方法使用其中的 bwa mem，最终获得癌症和正常样本的比对结果的bam格式文件 ，

所述步骤 B中， 采用 samtools软件提取 软件提取 read的位置和长度信息， HGSNV的 位点和覆盖该read数量信息， 其中使用 samtools view命令提取 read信息时， 使用 参数 -q 31过滤掉序列比 对质量（MAPQ）低于31的序列 ，其中 q表示过滤掉测序质量差的序列，同时使用 参数 -f 0x2 -F 0x18过滤掉未能正确匹配的read，其中 f表示提取符合一定要求的序列，F表示过滤符合一定要求的序列 ，使用 samtools mpileup命令提取 HGSNV信息时，使用参数 -q 20过滤掉序列比对质量低于20的序列，并 使用参数-Q 20过滤掉碱基质量小于20的序 列， 其中 Q表示过滤掉碱基质量差的序列 ；选取等位基因频率时，使用samtools mpileup的-l参数；使用该需要提前准备一个包含 SNP位点信息的 bed格式文件。

C1、将全基因组按照一定碱长度的 window为单位进行划分，对每个 window统计覆盖该 window的 read数量，统计时以每条 read的中点代表该 read的位置；

C2、对参考基因组创建索引文件，提高 GC含量的统计速度；对每一条染色体分别统计1、5、25、125个碱基间隔的区域内，鸟嘌呤（ G）和胞嘧啶（ C）的累积数量 ， 其中在统计某一个window中的 GC含量时，用 15 a\* 125 + b\*25 + c\*5 +d\*1的快速算法提取 ，其中 a,b,c,d表示系数变量 。

C3、以每个 window的 GC含量为自变，以每个 window的 read数量为因变量，拟合 read数量随 GC含量变化的函数；使用步骤 C1和步 骤 C2提取的各 window的 GC含量，通过如下弹性网络模型拟合 read数量随 GC含量变 化，其中使用 window的 GC含量 为变量x，使用 𝑥,𝑥2,𝑥3,𝑥4,𝑥5,𝑥6作为弹性网络模型 的输入变量，以 read数量为输出变，构建弹性网络模型如公式（14）所示：

𝐸(𝜆1，𝜆2，𝛽)=|𝑦−𝑋𝛽|2+ 𝜆2 + 𝜆1 ，𝜆1+𝜆2=1 （14）

公式（14）中， y表示 window内观测到的 read数量， X表示输入变量矩阵， 表示输入变量矩阵， 𝛽表示变量系数矩阵， j表示变量系数下标， P表示系数总， 𝜆1和𝜆2表示罚分系数。

C4使用拟合出的模型对全基因组 read数量进行调整。使用步骤 C3中的模型预测每一个 window理论上的 read数量 𝜇𝑔𝑐，基因组的平均GC含量定义为 𝜇，window内观测到的 read数量定义为𝑦， window内校正后的 read数量为 Y，那么校正公式如下（15） 所示：

𝑌= × 𝑦 (15)

步骤 E中， 使用步骤 D中 BIC-seq处理后的基因组片段为单位，计算所包含处理后的window数量， TRE的平均值以及方差，然后对片段的 TRE进行平滑化处理，方式如公式（16）所示， 针对每一个基因组片段，以 TRE的均值作为正态分布的均值𝜇，以 TRE的方差作为正态分布分布的方差 𝜎，计算出TRE在[𝜇−2𝜎,𝜇+2𝜎]范围内 window数量的分布，定义v为 TRE坐标，取值范围为 [𝜇−2𝜎,𝜇+2𝜎]，分辨率为 0.001，𝐶𝑤𝑖𝑛为该片段分配到 v位点的 window数量， 𝐶𝑇表示该片段内 window的总数，将所有片段window根据 TRE值平滑化后，可使片段内的window数量呈现正态分布，对所有片段各TRE位点对应的 window数求和汇总，得到基因组范围的 window随 TRE变化的分布：

(16)

步骤 F中， 以 0.001为分辨率，遍历 [0,1]范围内的所有P，使用类自回归模型，计算 Y(P)的值， Y(P)表现为多峰分布， 使用第二高峰内Y(𝑃)的最大值对应𝑃作为𝑃的计算结果，𝑀𝑡是 TRE的最大取值，这里将𝑀𝑡设置为3。

步骤 G中， 步骤 G包括3个步骤，步骤 G1中，遍历 [0,1]的 TRE区间作为 𝑋𝑓，过滤掉C(𝑋𝑓)小于 1000的TRE位点，计算使公式（5）取最大值时的 𝑋𝑓作为第一个实际观测 peak的均值。

步骤 I中，根据步骤 H计算的癌症样本纯度r和peak对应的拷贝数，计算peak的 MAF的期望值fb。

其中步骤 I包括 ：

I1，使用公式（17）计算 peak内 HGSNV的 MAF理论值：

公式（17）中，表示主要等位基因的拷贝数， 表示 peak的整体拷贝数， 由步骤 I得到， 𝑓表示该 peak内 MAF的理论值，可见当较大时， 𝑓有多种不同的可能值；

I2，利用负二项分布估计覆盖每个HGSNV位点的read总数的概率，使用公式（18） 计算负二项分布的概率 p和失败次数 r：

(17)

𝑝=1− ；𝑟= （18）

公式（18）中， m是 peak内所有 window中 read数量的均值，

v是 peak内所有 window中 read数量的方差，所求得p是用于负二项分布的随机变量成功概率，

r为随机变量 25 失败的次数，随机变量为覆盖某个 HGSNV中的 read数量；

I3，利用二项分布求得的覆盖某个HGSNV的 read数的概率 ，结合在一定read数量下， HGSNV只有两种基因型，服从二项分布规律利用公式（19）计算𝑓的校正值 𝑓𝑏，同一个 peak中，不同的𝐶𝑚𝑐𝑝计算得到不同的 𝑓𝑏，选择与该peak的 MAF观测均值最接近的𝑓𝑏作为该 peak的𝑓𝑏

(19)

公式（19）中， k表示在某个HGSNV位点，某一种等基因 A或 B的数量， d为覆 盖该 HGSNV的 read数量 ，r为随机变量失败的次数， p是用于负二项分布的随机变量成功的概率；

对每一个 𝑄𝑛，可推断获得基因组所有 peak对应的拷贝数和癌症样本纯度，从而每一个 peak求𝑓𝑏，进而得到所有 peak的 MAF的期望值得集合 的期望值得集合 {𝑓𝑏}。

1. 实验

我们将Accurity应用于172对TCGA肿瘤样品--正常样本。 其中61个样本的TRE直方图非常嘈杂，Accurity无法检测到有效值。 结果，Accurity成功获得了111个样本，其中有32个高覆盖率样本（覆盖率> 10），49个中等覆盖率样本（5-10个），30个低覆盖率样本（<5）。 如Aran等人报道的那样，我们对其进行安全性估计，并与ABSOLUTE，ESTIMATE和LUMP的估计进行了比较。（补充表格S1）。如果限制在Accurity成功的111个样本中， Accurity的总体性能（Spearman相关性ρ= 0.328 n = 111）（图5）与ABSOLUTE相当（ρ= 0.368 n = 153）。 大幅超越ESTIMATE和LUMP。

一旦覆盖率低于10，Accurity的安全性能就会下降。在高覆盖率（覆盖率> 10）样本（补充表格S2）中，安全性与ABSOLUTE相当，优于其他两种。在低覆盖率（<10）样本中，安全绩效下降。覆盖率为5-10，安全性（ρ= 0.187）与其他方法没有优势。在覆盖率低于5的情况下，Accurity的表现与覆盖5-10相似，略有改善（ρ= 0.291），与此同时，ABSOLUTE，ESTIMAT和LUMP表现都相当差。

我们使用更多模拟数据进一步证实了Accurity的性能对覆盖的依赖性（补充图S8）。一旦覆盖率低于10，安全性估计值与真实纯度水平的偏差增加到0.1以上。尽管安全性在覆盖率较低（<10）的样本中下降，但在低于5的覆盖率的样本中，其安全性仍然相对较好。这验证了我们的想法，即利用TRE直方图中的周期性来估计纯度。只要可以检测到TRE直方图中的明确片段，Accurity就可以产生合理的估计。

Accurity由主要由RUST实现，但也使用了大量C++代码以及一些python代码。经过测试，它可以运行在几乎所有Linux发行版。 从理论上讲，也可以在Windows和Mac平台上编译和运行，但我们还没有测试过。 对于5×肿瘤样品/正常样品的匹配对，Accurity的平均运行时间约为45分钟; 在Intel（R）Xeon（R）CPU E5-2670 v3 @ 2.30 GHz的单个内核上对于30×肿瘤样品/正常样品的匹配对运算约3小时; 最高RAM消耗低于4 GB。 由于其使用C ++ / RUST实现，在实际运行中，Accurity将比测试方法更快并占用更少的内存。

参考文献

1. Benjamini,Y. and Speed,T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res., 2012
2. Alkodsi,A, Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. Brief. Bioinf.,2015.
3. Andor,N.: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics, 2014.
4. Aran,D. Systematic pan-cancer analysis of tumour purity. Nat.Commun., 2015
5. Benjamini,Y. and Speed,T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res., 2012
6. Beroukhim,R. et al. The landscape of somatic copy-number alteration across human cancers. Nature, 2010
7. Bild,A.H. et al. Linking oncogenic pathways with therapeutic opportunities. Nat. Rev. Cancer, 2006
8. Boeva,V. et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. Bioinformatics,2014
9. Carter,S.L. et al. Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol., 2012
10. Cronin,M. and Ross,J.S. Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology.Biomark. Med., 2011
11. Degner,J.F. et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics, 2009
12. Elloumi,F. et al. Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med. Genomics, 2011
13. Favero,F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol, 2015
14. Garofalo,A. et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. Genome Med., 8, 79. 2016
15. Gusnanto,A. et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics, 28, 40–47. 2012
16. Hanahan,D. and Weinberg,R.A. Hallmarks of cancer: the next generation. Cell, 144, 646–674. 2011
17. Koboldt,D.C. et al. Var Scan somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res., 22,568–576，2012
18. Larson,N.B. and Fridley,B.L. Pur Bayes: estimating tumor cellularity and subclonality in next-generation sequencing data. Bioinformatics, 2013.
19. Li,Y. and Xie,X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. Bioinformatics, 2014.
20. Liu,B. et al. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. Oncotarget, 2013.
21. Mayrhofer,M. et al. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. Genome Biol., 2013.
22. Mwenifumbo,J.C. and Marra,M.A. Cancer genome-sequencing study design. Nat. Rev. Genet., 2013.
23. Navin,N. et al. Tumour evolution inferred by single-cell sequencing. Nature, 2011.
24. Oesper,L. et al. THet A: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol., 2013.
25. Potti,A. et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N. Engl. J. Med., 2006.
26. Ross,J.S. and Cronin,M. Whole cancer genome sequencing by next-generation methods. Am. J. Clin. Pathol., 2011.
27. Roychowdhury,S. and Chinnaiyan,A.M. Translating genomics for precision cancer medicine. Annu. Rev. Genomics Hum. Genet., 2014.
28. Sabbah,M. et al. Molecular signature and therapeutic perspective of the epithelial-to-mesenchymal transitions in epithelial cancers. Drug Resistance Updates Rev. Comment. Antimicrob. Anticancer Chemother.,2008.
29. Shah,S.P. et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature, 2012.
30. Su,X. et al. Purity Est: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics, 2012.
31. Wang,Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature, 2014.
32. Yadav,V.K. and De,S. An assessment of computational methods forestimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. Brief. Bioinf., 2015.
33. Yoshihara,K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun., 2013.
34. Yu,Z. et al. CLIm AT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. Bioinformatics, 30, 2014.
35. Zack,T.I. et al. Pan-cancer patterns of somatic copy number alteration. Nat. Genet., 2013.
36. Junttila M R, de Sauvage F J. Influence of tumour micro-environment heterogeneity on therapeutic response [J]. Nature, 2013.
37. Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer [J]. Nature Biotechnology, 2012, .
38. Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013.
39. Yuan Y, Failmezger H, Rueda O M, et al. Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors omplements enomic Profiling [J]. Science Translational Medicine, 2012.
40. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing [J]. Nature, 2011
41. Su X, Zhang L, Zhang J, et al. Purity Est: estimating purity of human tumor samples using next-generation sequencing data.[J]. Bioinformatics, 2012.
42. Larson N B. Pur Bayes: estimating tumor cellularity and subclonality in next-generation sequencing data [J]. Bioinformatics, 2013.
43. Gusnanto A, Wood H M, Pawitan Y, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data [J]. Bioinformatics, 2012.
44. Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013.
45. Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer [J]. Nature Biotechnology, 2012.
46. Li Y, Xie X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity [J]. Bioinformatics, 2015.
47. Mayrhofer M, Dilorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue [J]. Genome Biology, 2013.

在学取得成果

1. 在学期间所获的奖励

2015年11月，获得建龙奖学金

2015年11月，获得优秀三好学生荣誉称号

2016年11月，获得优秀学生干部荣誉称号

1. 在学期间发表的论文
2. 在学期间取得的科技成果

致 谢

我首先要感谢我的论文指导老师、北京科技大学计算机科学与技术系的田澍老师。田澍老师对我论文的研究方向做出了指导性的意见和推荐，在论文撰写过程中及时对我遇到的困难和疑惑给予悉心指点，提出了许多有益的改善性意见，投入了超多的心血和精力。田澍老师对我的帮忙和关怀表示诚挚的谢意!同时，还要感谢北京科技大学的授课老师们和所有同学们，大家在北京科技大学互相学习，互相帮忙，共同度过了一段完美难忘的时光。

此外，还要感谢朋友以及同学们在论文编写中带给的大力支持和帮忙，给我带来极大的启发。也要感谢参考文献中的作者们，透过他们的研究文章，使我对研究课题有了很好的出发点。

从开始接到论文题目到系统的实现，再到论文文章的完成，每走一步对我来说都是新的尝试与挑战，这也是我在大学期间独立完成的最大的项目。在这段时间里，我学到了很多知识也有很多感受，从一无所知，我开始了独立的学习和试验，查看相关的资料和书籍，让自己头脑中模糊的概念逐渐清晰，使自己十分稚嫩作品一步步完善起来，每一次改善都是我学习的收获，每一次试验的成功都会让我兴奋好一段时间。

我的论文作品不是很成熟，还有很多不足之处。但是这次做论文的经历使我终身受益。我感受到做论文是要真真正正用心去做的一件事情，是真正的自己学习的过程和研究的过程，没有学习就不可能有研究的潜力，没有自己的研究，就不会有所突破，那也就不叫论文了。期望这次的经历能让我在以后学习中激励我继续进步。