上海交通大学
2012级硕士研究生学位论文

利用生物信息学手段在急性髓系白血病中研究融合蛋白的致病机制以及筛选预后标志

硕士研究生：王焕威

学科专业：生物化学与分子生物学

导师：王侃侃

上海交通大学医学院附属瑞金医院

医学基因组学国家重点实验室

2015年10月28日

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

上海交通大学
学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，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，允许论文被查阅和借阅。本人授权上海交通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和汇编本学位论文。

保密□，在___年解密后适用本授权书。

本学位论文属于

不保密□。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘要

近年来随着基因芯片、第二代测序等高通量技术的发展，大量的生物数据被生产出来。利用生物信息分析这些数据，人们既可以对某一具体问题进行更为深入的机制研究，又可以从整体层面系统的了解生物运行网络，筛选出关键的分子标志。本课题就试图利用生物信息学的手段在急性髓系白血病（acute myeloid leukemia，AML）中对一种常见的融合蛋白 AML1/ETO 的致病机制进行深入研究，并系统整合基因、小 RNA 和 DNA 甲基化三维数据筛选出预后意义的分子标志。第一部分我们利用 ChIP-seq 技术获得了高可信度的 AML1/ETO 和野生型 AML1 蛋白的靶位点，通过下游的生物信息学分析发现野生型 AML1 参与到 AML1/ETO 的致病过程中。具体包括 AML1 蛋白与 AML1/ETO 分别结合在临近但不同的 motif 位点上，与 AML1/ETO 形成 complex，同时与 AML1/ETO 的激活和抑制作用有关，与 AML1/ETO 的结合信号强弱决定了 AML1/ETO 转录方向。同时，我们还预测 AML1/ETO 的激活作用与招募共激活因子 AP-1 蛋白质有关。第二部分通过整合急性髓系白血病 TCGA 的基因表达数据、小 RNA 表达数据、基因的 DNA 甲基化水平，以及病人的临床资料和生存情况，系统挖掘在急性髓系白血病发病过程中对病人预后有影响的分子标志。最终形成整合成一个集合，对急性髓系白血病病人的预后情况起到预测作用。

关键词：生物信息学、ChIP-seq 技术、急性髓系白血病、AML1/ETO 融合蛋白、预后标志物

Abstract

More and more biologic data has been produced recently with the development of microarray, second-generation sequencing and other high-throughput technologies. To analysis these data, bioinformatic methods become important both in a particular gene/protein research and in a system network study. This project aims to doing bioinformatics analysis to study the pathogenic mechanism of the AML1/ETO fusion protein, and to integrate gene expression, microRNA expression and DNA methylation data to identify the prognosis biomarkers in acute myeloid leukemia (AML). In the first part, we identify high confidence AML1/ETO and wild-type AML1 binding sites using chromatin immunoprecipitation sequencing and find the genome-wide interplay between AML1/ETO and wild-type AML1. We demonstrate that AML1 and AML1/ETO preferentially bind to adjacent and distinct short and long AML1 motifs on the co-localized regions, AML1 exists in the AML1/ETO complex and the binding signals between AML1/ETO and AML1 determine the direction of AML1/ETO transcriptional regulation. In addition, we predict that AML1/ETO transactivated gene expression through recruiting the co-factor AP-1. In the second part, we identify the significant genes, microRNA and DNA methylation associated with prognosis using TCGA AML data. And we report a combined signature to predict the prognosis of AML patients.

Keywords: Bioinformatics, ChIP-seq, acute myeloid leukemia, AML1/ETO fusion protein, prognosis biomarker

目录

引言

生命科学是现代科学中至关重要的一门学科，而生命科学的研究往往离不开对生物数据中所蕴含的重要信息的挖掘，即生物信息学。近年来，随着基因组学和高通量技术的发展，人们面临着海量的生物数据，生物学研究进入了大数据的时代。因此，如何挖掘其中的生物学意义，进一步分析其中的重要信息就愈发重要[1]。

生物信息学是一门集数学，计算机科学和生物学的工具以及技术于一体的涵盖了生物信息的获取、处理、存储、分配、分析和阐述等各个方面以理解海量的生物学数据为目的的学科。生物信息学在医学遗传学研究领域具有广泛的应用，其研究的范围涉及基因组学、蛋白质组学、代谢组学、药物设计、调控网络、分子进化、比较基因组学、系统生物学等。作为一门交叉学科，生物信息学所涉及的学科也是相当之多的，如涉及统计学、概率论等数学学科，编程、算法、机器学习、数据库建立与查询等计算机学科。拿编程语言来说，就会涉及 C、C++、Java 等高级语言，perl、python 等脚本语言，SQL、XML 等数据库语言，matlab、R 等用于统计和画图的编程语言。

虽然生物信息学是随着基因组学和高通量技术而引起人们的重视，成为近十年来最为火热的学科之一，但其实它已经有相当长时间的历史，其历史甚至要早于“生物信息学”一词的产生。“生物信息学”一词（即 bioinformatics）是于 1970 年 Ben Hesper 和 Paulien Hogeweg[2] 率先提出，但是在 1965 年，美国科学家 Zuckerkandl 和 Pauling[3] 首次提出用分子序列来进行进化方面的研究，这被认为是生物信息早期非常重要的工作。科学家 Ouzounis 和 Valencia[4] 综述了他们个人因为非常重要的 20 件生物信息领域的大事件，其中包括蛋白质折叠的模拟、蛋白质序列的预测等等。

近年来生物信息的火热得益于基因组学和高通量技术的发展（如图 1A），1973 年的 Maxam-Gibert 测序和 1975 年提出的 Sanger 测序是第一代测序技术；1995 年，Science 杂志发表连续发表两篇 microarray 的文章[5,6]，使得人们第一次可以高通量的研究整个基因组的表达情况（即转录组）。而到 2006 左右，由 Roche 公司推出的 454

测序仪、Illumina 公司推出的 Solexa 测序仪以及 ABI SOLID 测序仪为代表的二代测序技术的推出，成为了推动基因组学研究最主要的功臣，其“边合成边测序”的原理使得高通量测序变成了显示。目前测序技术不断发展，Illumina Miseq、Illumina Hiseq、Ion PGM 和 Illumina X Ten 等测序仪使得测序变得更快、更便宜、更准确。2008 年，以 Heilscope 和 Nanopore 为代表的三代测序也已经推出，与二代测序相比，尽管目前还存在错误率过高等问题，但是其对长序列的测序能力和单分子测序的原理使人们对未来有了更多的期待[7]。测序技术的发展和一系列相关计划的实施，使得生物数据的产生呈现了“爆炸式”的增长。根据美国国家生物技术信息中心的 DNA 序列的大型数据库 GenBank 的数据显示（如图 2），科学界产生的 DNA 序列的数据在最近十年有了一个高速的增长，而且这种增长的速率直到今天都没有下降。因此，面对测序技术快速的发展和生物数的增加，生物信息的分析就变得不可或缺。

由于众多生物性状和疾病与基因组密切相关，人们对基因组进行了大规模的重测序，这其中包括全基因组测序（Whole Genome Sequencing）和全外显子测序（Whole Exon Sequencing）。2002 年启动了人类单倍体型图计划（HapMap 计划）用于构建人类基因组中常见遗传多态位点目录，从而将遗传多态位点和特定疾病风险联系起来。2008 年启动的千人基因组计划（1000 Genomes 计划）鉴定除了更多的多态位点，并且鉴定出更长、更复杂的结构变异。而 2005 年启动的美国癌症基因组计划（TCGA 计划）和 2007 年启动的国际癌症基因组计划将大规模测量癌症病人的基因组，找到与癌症发病和治疗相关的基因组变异，使得基因组信息更好地应用到临床诊断和抗肿瘤药物开发的实践当中。

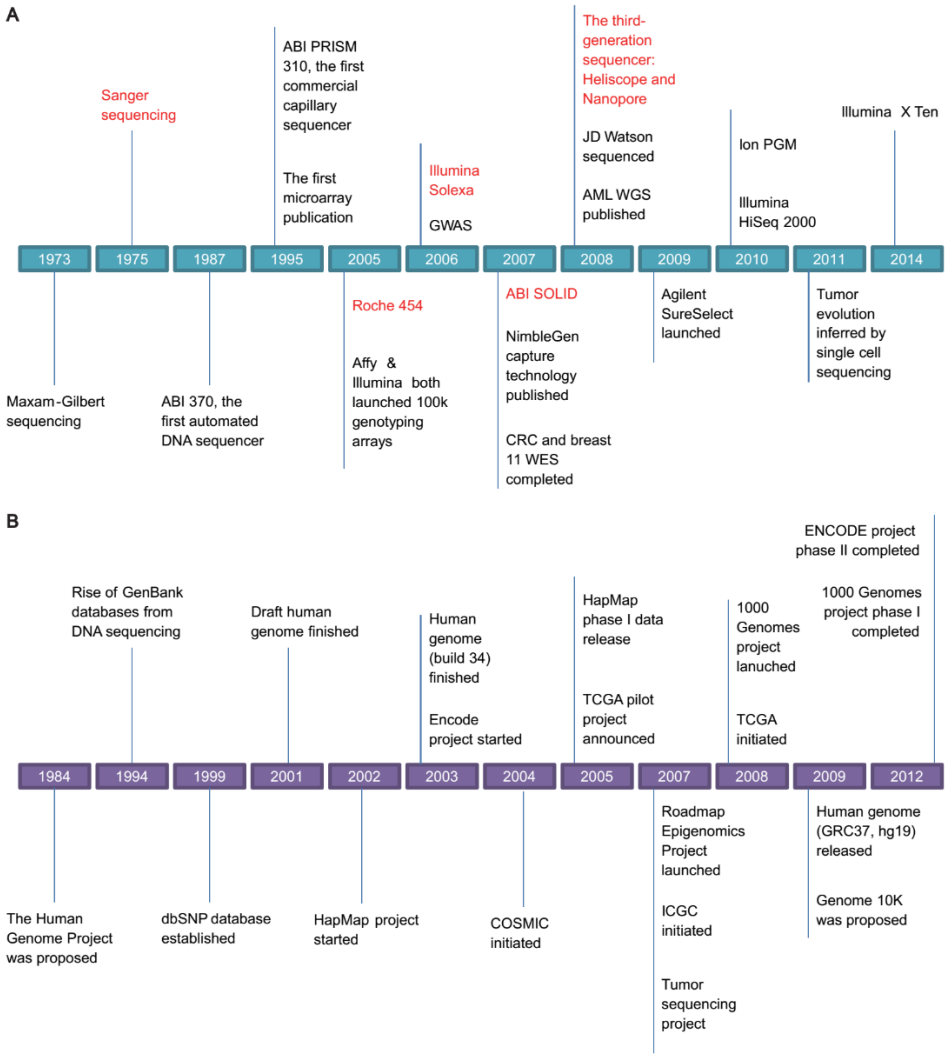


图 1 DNA 测序技术（A）及相关计划（B）的时间轴

Figure 1. The timeline of DNA sequencing technology (A) and related projects (B)

除了单纯对基因组的序列进行重测序外，更多的功能基因组测序得以实现，如研究转录组的 RNA-seq 技术、研究蛋白质与 DNA 相互作用的 ChIP-seq 技术、研究 DNA 甲基化的 MeDIP-seq 和 Bisulfite-seq 技术、研究 3D 基因组的 ChIA-PET 和 Hi-C 等技术。而由于目前测序平台的样本都是一群细胞，为避免细胞异质性，单细胞测序技术也被开发出来。2007 年启动 2012 年完成的 DNA 元件百科全书计划（ENCODE 计划）就解析了人类基因组的调控信息，更新了人们对于功能基因组的认识。而 Roadmap 计划则是利用新一代测序对于人类的表观基因组进行了全景图式的研究。

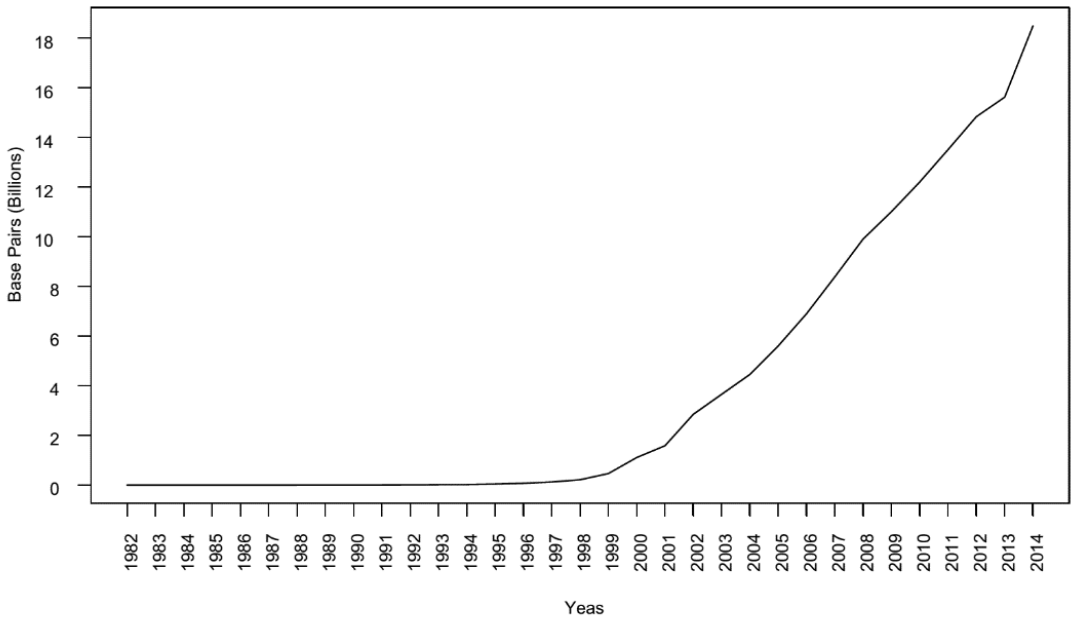


图 2 GenBank 数据库近年来 DNA 序列的增长；

数据来源：ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt，2015 年 2 月 2 日下载

Figure 2. The growth of sequenced DNA base pairs in GenBank Database

Data source: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt, downloaded in Feb 2nd, 2015

本人研究生期间在上海市瑞金医院血液学研究所和国家基因组学重点实验室学习工作，利用生物信息学手段来处理、分析各种急性髓系白血病（AML）的数据。这份硕士学位论文主要包括了其中两部分的内容，第一部分是利用研究蛋白质与 DNA 相互作用的高通量技术 ChIP-seq 来研究融合蛋白 AML1/ETO 的致病机制，尤其是其与野生型蛋白 AML1 之间动态平衡的关系，第二部分是整合多组学数据和临床资料，寻找和评估预后标志。下面我将详细描述。

第一部分 利用 ChIP-seq 技术研究 AML1/ETO 融合蛋白及野生型 AML1 蛋白致病机制

1、 绪论

一个生命体的不同组织不同器官的不同细胞虽然具有一样的基因组但表现出的功能却千差万别，这往往是由于不同细胞基因表达的不同所造成的。转录因子是一类可以调控细胞基因表达（激活或者抑制）的蛋白质，它们可以特异性地识别某段 DNA 序列，直接结合在基因组的转录调控元件上，与其他的共调节因子相互作用，对近端或者远端的基因表达进行转录调控 [8]。

染色质免疫沉淀-测序技术（ChIP-seq 技术）是一种基于新一代测序技术发展起来的，在全基因组水平上研究转录因子（或其他蛋白质）与 DNA 相互作用的高通量技术 [9]。通过特异性识别转录因子的抗体将转录因子以及与其结合的 DNA 片段富集出来，通过高通量测序技术和生物信息学分析，可以鉴定出该转录因子在全基因组水平上的结合位点。较之前的单基因或单位点的分子生物学实验技术和基于芯片的 ChIP-on-chip 技术，ChIP-seq 技术具有高分辨率、低噪音和高覆盖度等优势。

随着测序技术的发展和生物大数据时代的到来，生物信息学越来越被人们所重视。ChIP-seq 生物信息学的基本分析主要可以分为序列比对、peak calling 和 motif 分析等步骤 [10]。我们可以将 ChIP-seq 的测序片段回帖到参考基因组上从而鉴定出转录因子结合的相应位置，利用统计模型和相关软件对转录因子结合区域、结合分步、结合信号进行定量，通过结合区域的序列特征进行 motif 分析 [10]。

进一步，人们利用生物信息学还可以对数据进行更精细的挖掘。例如，细致地分析 motif 可以鉴定出同一个转录因子在不同位点上 motif 的细微差别。我们课题组之前就通过转录因子 PU.1 的 ChIP-seq 数据鉴定出两种 PU.1 的 motif，并且发现它们参与到了不同的生物学功能 [11]。其次，通过 motif 分析还可以预测其他转录因子是否也结合在相同的位置。有文章报道通过 ChIP-seq 数据的 motif 分析，发现 AML1/ETO 蛋白可以和 HEB、LMO2 等 E-box 家族的蛋白相互作用 [12]。另外，结合上下游基因

了解转录因子在 DNA 上的结合对相关基因的调控作用。人们利用 Myc 与 Miz 的 ChIP-seq 数据，发现 Myc 与 Miz 的 ChIP-seq 信号强弱可以决定 Myc 在 Myc 诱导的癌症细胞中对于下游基因调控的方向 [13]。最后，利用 ENCODE 等公共数据库中的 ChIP-seq 数据，人们可以在转录因子结合区域富集共转录调控因子等 [14,15]。这一系列的生物信息手段为我们更深入全面地研究转录因子在细胞中的功能提供了可能。

AML1 蛋白（又称 RUNX1、CEBFα2 或者 PEBP2αB）是在正常胎儿和成人造血过程中均起到重要作用的转录因子，对血细胞的增值和分化进行转录调控 [16]。AML1 在白血病中涉及到多种染色体异常 [17-19]，其中 AML1/ETO 是急性髓系白血病（AML）中最为常见的融合蛋白之一 [19]。AML1/ETO 可以促进细胞自我更新，抑制细胞分化，其诱导白血病生成已经被多种条件性转基因或移植动物模型所证实 [20]。t(8;21)所形成的 AML1/ETO 融合蛋白由包含 DNA 结合结构域（RHD 结构域）的 AML1 的 N 端部分和包含 NHR1-4 结构域的 ETO 蛋白的 C 端部分组成。传统观点认为，AML1/ETO 因具有与 AML1 相同的 DNA 结合结构域而结合在 AML1 的位点，通过原 ETO 蛋白的 NHR1-4 结构域招募 HEB、N-CoR/SMRT[21]、mSin3a、HDACs[22] 等共抑制因子来抑制 AML1 靶基因的表达，对野生型的 AML1 蛋白起到抑制作用。另外，由在细胞中 t(8;21)易位仅发生在一条染色体上，因此另一条染色体会依旧会表达野生型的 AML1 蛋白。最近越来越多的研究发现野生型的 AML1 蛋白也对白血病的发展起到关键作用，尤其是与 AML1/ETO 相关的白血病 [23,24]。

为了进一步研究 AML1/ETO 蛋白质的致病机制，尤其是与野生型 AML1 蛋白之间的关系，我们利用 ChIP-seq 技术检测它们在 t(8;21)型白血病细胞中的结合位点，并发现他们结合在相邻的位点，倾向结合于相似但是不同的 motif 上，并可以形成复合体。而且 AML1/ETO 的转录调控与 AML1/ETO 和 AML1 结合强弱个 AML1/ETO 蛋白招募共激活因子 AP-1 相关。

2、 材料与方法

2.1 ChIP-seq 的数据分析

ChIP-seq 的数据分析主要分为：序列比对，peak calling 等步骤，以下将会分别描述。同时，我将这些步骤用 Python 语言进行了整合，具体代码参见附录 1 和 2。

2.1.1 序列比对

Illumian 所测的原始的 ChIP-seq 数据下机后，经过自带的 base calling 程序可以转换成 FASTQ 文件。之后，我们需要将 FASTQ 文件比对到参考基因组上，该步骤成为序列比对，可以通过 BWA[25]、Bowtie2[26] 等软件完成。这里我们采用 BWA 软件将 FASTQ 文件比对到参考基因组（NCBI Build 37，hg19）上去。具体流程如下：

1、 在 UCSC 网站上下载参考基因组 hg19 的原始序列 FASTA 文件（http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz）。由于 UCSC 网站上的参考基因组是每一条染色体单独存储的，故解压之后利用“cat”命令，将多个文件合并，并命名为“hg19.fa”。

2、 由于人类基因组有 30 亿对碱基，因此比对软件往往需要首先对参考基因组的序列建立索引。对于 BWA 软件建立索引，我们采用“bwa index -a bwtsm hg19.fa”命令。

3、 BWA 软件有两个命令（“bwa aln”和“bwa men”）可以用来进行序列比对。“bwa aln”比较适合序列比较短且测序量不是很大的数据，而“bwa men”比较适合于序列比较长且测序量很大的数据，尤其是全基因组数据。由于我们的 ChIP-seq 数据，测序长度为 35bp，单端，且每个样本的测序片段在 2 千万（20M）以内，故我们采用“bwa aln”命令。

4、 BWA 比对之后输出的文件为 SAM 文件。该文件格式不仅可以无损的保存原始信息（如片段 ID，测序序列和测序质量）还可以存储测序序列的位置等信息。另外，我们还可以利用软件 samtools[27] 将 SAM 文件中比对之后的序列进行排序（“samtools sort”命令），将 SAM 文件转换为更小的二进制 BAM 文件（“samtools view –bS”命令），对 BAM 文件进行索引生成 BAM.BAI 文件（“samtools index”），还可以对比对结果进行简单的统计（“samtools idxstats”

和“samtools flagstat”）。

5、 若一条测序序列比对到基因组上的多个不同位置，其也会影响下游分析的结果。BWA 软件会把这种测序序列的比对质量分数赋值为 0。故用“samtools view –bq 1”命令，删除比对到多个位置的序列和比对质量分数小于 1 的序列。

6、 由于 ChIP-seq 在测序过程中，会对 DNA 片段进行 PCR 扩增，故在分析时需要用“samtools rmdup”命令将重复的序列（即相同的序列片段在文件中多次出现）删除。

利用上边的流程，我处理了实验室自己产生的 6 套 ChIP-seq 数据（2 套为对照），具体比对情况如下：

表 1 ChIP-seq 序列比对结果统计

Table 1. Overview of ChIP-seq alignment

ChIP-seq	Cell line	Tags	Mapped tags	Uniquely mapped tags	Non-redundant tags
Anti-AML1 (C19)	Kasumi-1	18,055,494	15,993,109	13,247,510	13,062,289
Anti-AML1 (N20)	Kasumi-1	18,368,604	16,181,828	13,450,563	13,183,554
Anti-ETO	Kasumi-1	16,440,220	15,073,523	12,458,712	12,248,193
Total	Kasumi-1	18,065,658	16,423,856	13,276,162	13,078,211
Anti-AML1 (C19)	U937	22,169,579	20,935,858	17,590,487	17,279,866
Total	U937	16,835,674	16,835,674	15,969,724	12,877,278

另外，我们的 ChIP-seq 数据已经上传到 GEO 网站上，编号为 GSE65427。

2.1.2 Peak calling 及相关分析

将原始的 FASTQ 进行比对得到 SAM/BAM 文件之后，下面我们要对序列富集的区域（peak）进行鉴定，即目标蛋白在全基因组上的结合位点。该步骤可以使用 SPP[28]、

MACS[29]、MACS2、HPeak[30]等多个软件来完成，这里我们采用广泛使用的 MACS 软件（版本号 1.4.2）。而鉴定完 peak 之后，我们还可以进行很多相关的分析，如这些 peak 在全基因组上位置的分布，这些 peak 所对到的基因，多套 ChIP-seq 数据鉴定出的 peaks 之间有多少重合，提取 ChIP-seq 在特定位置的信号信息进行热图展示等。具体流程如下：

1、 利用软件 MACS 寻找序列富集的区域 peak。为了得到更准确的结果，每一个细胞系的 ChIP-seq 数据均有一套 Total 数据作为背景进行校正。对于 Anti-AML1 (N20)的 ChIP-seq 数据，p-value 设定为 10E-5，其他数据的 p-value 设定为 10E-8。同时为了直观的展示 ChIP-seq 数据的信号，我们设置了“-w --space=10”参数来生成 wig 文件，之后 wig 文件可以利用 UCSC 的 wigTobigWig 工具转换成 bigWig 文件，导入 IGV 中进行可视化。另外，由于是数据为人类基因组，“-g”参数我们设置为“hs”。最终，在 MACS 输出文件中我们主要使用的是 XLS 文件（储存了 peak 的位置、区域、最高峰（summit）的位置、p-value、FDR 值等信息）和 wig 文件（储存了在全基因组上各个位置峰的信号高低）。

2、 利用 BEDtools[31] closest 工具进行 peak 在全基因组上的位置分布分析并与最近的基因相关联。Peak 在全基因组上的位置分布，主要指的是该 peak 到底是在全基因组上基因的启动子区域、外显子区域、内含子区域、上游调控区域、下游调控区域还是基因间区域。故，我们用 UCSC 的 Table Browser 工具下载了 Refseq 的基因注释文件（http://genome.ucsc.edu/cgi-bin/hgTables）。利用 BEDtools closest 工具将这些 peak 与离它们最近基因相关联，之后再判断该 peak 是出于该基因的哪个区域。不同区域的定义如下：启动子区域为基因转录起始位点上游 3kb 到下游 1kb，外显子和内含子区域由 Refseq 数据库定义，上游调控区域为转录起始位点上游 3kb 到上游 20kb，下游调控区域为转录终止位点到下游 20kb，其他区域为基因间区域。同时，我们认为如果一个 peak 不是处在基因间区域，就可以与该基因进行关联。

3、 当处理完多套 ChIP-seq 数据，鉴定了各自的 peak 之后，我们就可以分析不

同数据之间重叠的部分所占的比例为多少。该步骤可以采用 BEDtools intersect 工具分析，利用 R 语言中的 VennDiagram 包画文氏图进行展示。

4、 在 ChIP-seq 的分析过程中，常常需要对于基因组上特定位置（peak 区域、基因区域、summit 左右 1kb 等）的数据信号进行提取，用来进行展示（热图）或者分析（累计线图，aggregation plot）。这个过程我们是利用 UCSC 的 bigWigSummary 工具完成的。由于，还需要根据测序的深度、特定区域的宽度等对信号值进行校正，所以我们写了一个 Python 脚本来完成这个步骤（见附录 3）。

除了我们自己实验室的 ChIP-seq 数据，我们还分析了很多已经发表了的公共的 ChIP-seq 数据[12,23,32-35]（表 2），其流程与上边一致。

表 2 论文中所用的公共的 ChIP-seq 数据

Table 2. Published ChIP-seq data used in this paper

Source	Antibody	Cell line and treated method	Reference
GSM1113427	AML1	Kasumi-1	23
GSM1113428	AML1	Kasumi-1	23
GSM722704	AML1	Kasumi-1	33
GSM850824	AML1	Kasumi-1 cells transfected with AML1/ETO siRNA	33
GSM850823	AML1	Kasumi-1 cells transfected with mismatch AML1/ETO siRNA	33
GSM726978	AML1	SKNO-1	32
GSM610330	AML1	CMK	34
GSM837994	AML1	CCRF-CEM	35
GSM1082306	AML1/ETO	Kasumi-1	12
GSM1113429	AML1/ETO	Kasumi-1	23

GSM1113430	AML1/ETO	Kasumi-1	23
GSM722705	AML1/ETO	Kasumi-1	33
GSM585588	AML1/ETO	Kasumi-1	32
GSM585589	AML1/ETO	SKNO-1	32
GSM1082309	E2A	Kasumi-1	12
GSM1082308	HEB	Kasumi-1	12
GSM1082311	LMO2	Kasumi-1	12

2.2 motif 分析

所谓 motif 就是指重复性的某种模式，例如序列 motif、结构 motif、网络 motif 等等。我们这里讲的是 DNA 序列 motif，它是指短的、重复性的、被认为有生物学功能的 DNA 序列模式 [36]。motif 的研究方法有很多，早期人们主要利用实验的方法（如 DNase footprinting，gel-shift or reporter construct assays 或者 SELEX 等）。最近，由于 ChIP-on-chip 和 ChIP-seq 等高通量数据的产生，人们开始利用计算机对 motif 进行分析。

Motif 的表示方法一般有两种，一种是共识序列（Consensus），一种是矩阵的方法。所谓共识序列指的是一串字符串，利用 IUPAC code（International Union of Pure and Applied Chemistry）的规定，用字母 A、G、C、T 代表序列腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶，另外还用其他字母代表 2 个或者多种可能性（如用 N 代表该位置可能是四种碱基中的任意一个，用 R 代表该位置可能是 A 或者是 G）。而矩阵的方法表示 motif 又可以分为计数矩阵（count-matrix），位置频数矩阵（position frequence matrix，PFM）和位置权重矩阵（position weight scoring matrix，PWM）。

利用 ChIP-seq 的数据进行 motif 分析，主要有这么几种：（1）motif 发现（motif discovery），一般指的是输入一些序列，在其中找到一个或者多个 motif，该步骤可以用 MEME、MEME-ChIP[37]、DREME[38]、AMD[39]、HOMER[40] 等软件；（2）motif 注释（motif annotation），是指对于找到的一个 motif，在已知的 motif 的数据库（如 TRANSFAC[41] 和 JASPAR 等）找到相似的 motif，进而知道该 motif 是属于哪一个蛋

白的，该步骤可以用 TOMTOM[42] 等软件；（3）motif 聚类（motif clustering），是指对于多个 motif，进行两两的相似性比较后，聚类出哪些 motif 可以归为一类，哪些 motif 不能，该步骤可以使用 STAMP[43] 等工具；（4）motif 扫描（motif scanning 或者 motif search），是在一些序列中找到 motif 具体所处的位置，该步骤可以使用 FIMO[44] 等软件；（5）motif 还有很多其他的分析，如 motif 的功能分析（如软件 GOMO），motif 中心富集分析（如软件 CentriMo）等等。本研究采用了前四种分析，即 motif 发现分析、motif 注释分析、motif 聚类分析和 motif 扫描。具体步骤如下：

1、 Motif 发现分析：为了确定 AML1/ETO 和 AML1 重叠的 peak 区域、AML1/ETO 单独的 peak 区域、AML1 单独的 peak 区域有哪些 motif，我们使用了三个从头 motif 发现软件，分别为 AMD[39]、MEME-ChIP[37] 以及 HOMER[40]，对 AML1/ETO 或者 AML1 的最高点的左右两边 150bp 的区域进行扫描（详见附录 4）。所有的参数均为默认参数，除了 HOMER 软件的参数“-len”参数设置为了“-len 6,7,8,9,10,11,12”以扫到更多其他长度的 motif（“-len”是用来设置扫到 motif 的长度，默认的设置为“-len 8,10,12”；而超过 12bp 的 motif 会使得该软件超出系统内存）。

2、 Motif 注释：在从头找到了一些 motif 之后，我们利用软件套件 MEME 中的工具 TOMTOM[42] 将找到的 motif 与 TRANSFAC 数据库 [41]（版本号 9.2）中包含的已知的 motif 进行比较，注释这些新找到的 motif。

3、 Motif 聚类：我们是采用网络工具 STAMP[43]（www.benoslab.pitt.edu/stamp/）完成的，利用了 KL 散度（Kullback-Leibler divergence）对相似性进行比较。

4、 Motif 扫描：我们采用的是 MEME 软件套件中的工具 FIMO[44] 在 AML1/ETO 和 AML1 重合的 peak 上对我们关注的长 AML1 motif 和短 AML1 motif 进行的扫描，p-value 设置为 5E-4。在分析长短 AML1 motif 的时候，当两个 motifs 有重合的时候，我们会保留得分高的一个。而对于长短 AML1 motif 在 AML1/ETO 和 AML1 重合区域的频率分布只计算了同时含有长的和短的 motif 的 peak。对于一个 peak 上有多个长 AML1 motif 或者短 AML1 motif 的情况，我们只保留了离 summit 最近的长 AML1 motif 或者是短 AML1 motif。

2.3 RNA-seq 和表达谱数据分析流程

为了研究 AML1/ETO 对于基因表达的影响和野生型 AML1 如何参与 AML1/ETO 的转录调控，我们下载了三套公共的基因表达数据：一套为 Kasumi-1 细胞中 AML1/ETO 敲除前后的 RNA-seq 数据（GSE43834[12]），一套为 SKNO-1 细胞中的 AML1/ETO 敲除前后的表达谱数据（GSE34594[33]），一套为 AML1/ETO 阳性或是阴性的 AML M2 型病人的表达谱数据（GSE14468[45]）。以下为 RNA-seq 和表达谱数据具体的分析流程。

2.3.1 RNA-seq 分析

与 ChIP-seq 的分析类似，RNA-seq 的分析首先也要进行序列的比对。但是，由于 mRNA 的形成经过了剪切，去掉了内含子，因此测序的 RNA 片段或者 cDNA 片段可能在基因组上是两段或者几段。我们采用目前比较普遍使用的 RNA-seq 比对软件 TopHat[46]（版本号 2.17.0），先根据基因注释参考文件 GTF 文件在参考基因组 FASTA 上构建出转录组的序列，将 RNA-seq 片段先比对到转录组，然后把没有比对上的 RNA-seq 片段再比对到基因组上。我们的分析依然采用的是 Refseq 基因注释数据库和 hg19 参考基因组（详见“ChIP-seq 的数据分析”部分）。为了加快比对的速度，我们设置了参数“--no-coverage-search”，其他参数均为默认设置。

对于 RNA-seq 获得差异表达基因有多种方法和软件，如 Cufflinks，HTseq-count+DEseq，HTseq-count+edgeR，Gfold 等等。我们采用是软件 Gfold[47]，其算法对于不含有生物重复的 RNA-seq 数据具有相当的优势。具体说来，它包含两个部分“gfold count”和“gfold diff”分别用来计算基因上 RNA-seq 的片段数和找到差异表达基因。结果会对于基因注释数据库中的每个基因给出 RPKM 值（Reads per kilobase er Million mapped reads）、变化倍数（fold change）、Gfold 值（Generalized fold change，比 fold change 更准确的统计值）。之后，我们根据敲出 AML1/ETO 前后的 RPKM 值的和大于 1 且 Gfold 值的绝对值大于 1 的原则，挑选出了 Kasumi-1 细胞中 AML1/ETO 激活和抑制的基因（表 3）。

2.3.2 表达谱数据分析

两套表达谱的数据（GSE34594[33] 和 GSE14468[45]）采用的均是 Affymetrix HGU 133

plus 2.0 的平台。首先，我们对原始的 CEL 文件利用 RMA（Robust Multi-array Average）算法[48]进行标准化处理。该步骤我们使用的是 R 语言中的软件包 affy 中的 rma()函数完成，相对应的芯片环境 CDF 文件会自动加载，结果输出每个探针在每个样本中的表达值。注意为了方便运算，这个表达值是取过 log2 后的结果。然后，我们会对每个探针所对应的基因进行注释，去掉那些对不到基因的探针。另外，如果一个基因有多个探针对应，我们会选择表达值最高的那个探针进行后续分析。该流程也已经利用 Python 语言进行了整合，详见附录 5。

对于 SNKO-1 的表达谱数据（GSE34594[33]），我们根据至少两个时间点（共四个时间点）有超过 1.5 倍调变的标准进行筛选。对于 AML1 M2 型病人的表达谱数据（GSE14468[45]），我们利用 SAM（significance analysis of microarrays）算法[49]进行分析，并根据 FDR 为 0 且表达变化大于 1.5 倍的标准进行差异表达基因的筛选（表 3）。

表 3 论文中所用到的基因表达数据

Table 3. Gene expression data used in this study

Gene set name	Description	Source	Method	Reference
Kasumi-1_AML1/ETO_down	Genes upregulated after AML1/ETO knockdown in Kasumi-1 cells	GSE43834	RNA-seq	12
Kasumi-1_AML1/ETO_up	Genes downregulated after AML1/ETO knockdown in Kasumi-1 cells	GSE43834	RNA-seq	12
SKNO-1_AML1/ETO_down	Genes upregulated after AML1/ETO knockdown in SKNO-1 cells	GSE34594	Microarray	33
SKNO-1_AML1/ETO_up	Genes downregulated after AML1/ETO knockdown in SKNO1 cells	GSE34594	Microarray	33

AML-M2_AML1/ETO_down	Genes upregulated in t(8;21) positive AML-M2, as compared with t(8;21) negative AML-M2	GSE14468	Microarray	45
AML-M2_AML1/ETO_up	Genes downregulated in t(8;21) positive AML-M2, as compared with t(8;21) negative AML-M2	GSE14468	Microarray	45

2.4 Gene Set 分析

一个基因集合被定义为一组在特定条件下具有相似生物学特征的基因，例如 AML1/ETO 上调或者下调的基因。富集的显著性采用二项式检验进行评估，并计算 Z score。详细的描述可见之前我们课题组发表的文章[50]。

2.5 GSEA 分析

GSEA[51]（Gene Set Enrichment Analysis）分析是一种用来确定是否在两个条件下基因表达量的变化与已经定义了的一组基因集合相关的分析方法。这里我们利用 GSEA 的算法，去评估是否 AML1/ETO 调变的基因与 AML1/ETO 和 AML1 的结合强弱相关联。输入的数据为：（1）基因集合：AML1/ETO 激活或抑制的基因；（2）根据 AML1/ETO 与 AML1 结合信号的比值排过序的 AML1/ETO 和 AML1 共同的结合位点；（3）所有 AML1/ETO 和 AML1 共同结合位点以及所对应的基因。

另，Gene Set 分析和 GSEA 分析基本思路类似。但 Gene Set 分析的输入数据是两个基因集合，从而判断他们直接的富集程度，而 GSEA 分析的输入数据是一个基因集合和全基因组上基因的调变排序。另外，其他方面也有一些差别，如利用的统计学方法等。

2.6 （共）转录调控因子的富集分析

为了寻找潜在的（共）转录调节因子，我们首先用 Kasumi-1 细胞系中 AML1/ETO 敲出前后的基因表达数据（GSE43824[12]）鉴定出 AML1/ETO 调变的基因列表，即 AML1/ETO 敲除前后的差异表达基因。之后，我们再筛选出其中被 AML1/ETO 和

AML1 都结合的基因。基于这些基因，我们进行了两个独立的分析：一个是 ENCODE ChIP-seq Significance Tool[14]（http://encodeqt.simple-encode.org），另一个是 Cscan[15]（http://159.149.160.51/cscan/）。这两个工具均收集了上百套已经发表的 ChIP-seq 数据，根据用户输入的基因列表在各自的数据中寻找潜在的转录调节因子。另外，这两个工具均采用超几何分布进行富集显著性评估，并利用 Benjamini and Hochberg 方法进行多重假设检验校正。

3、 结果

3.1 通过 ChIP-seq 数据鉴定出高可信度的 AML1/ETO 和 AML1 的结合位点

为了研究 AML1/ETO 在 t(8;21) AML 中的致病机制及与野生型 AML1 之间的关系，我们在 Kasumi-1 细胞系中利用了 ChIP-seq 技术来鉴定它们在全基因组上的结合情况。由于 Kasumi-1 细胞系表达 AML1/ETO 和 AML1 但是不表达 ETO 蛋白，于是我们用抗 AML1 蛋白 C 端的抗体 Anti-AML1 (C19)来结合野生型的 AML1 蛋白，用抗 ETO 蛋白的抗体来结合 AML1/ETO 蛋白，用抗 AML1 蛋白 N 端的抗体来同时结合 AML1/ETO 和 AML1 蛋白（如图 3A 和 B）。之后，我们将 Anti-AML1 (C19) ChIP-seq 所鉴定的 peak 与 Anti-AML1 (N20) ChIP-seq 所鉴定的 peak 相交获得高可信度的 AML1 的结合位点，将 Anti-ETO ChIP-seq 所鉴定的 peak 与 Anti-AML1 (N20) ChIP-seq 所鉴定的 peak 相交获得高可信度的 AML1/ETO 的结合位点。如图 3C 所示，我们找到了 16,182 个高可信度的 AML1 的结合位点和 14,548 个高可信度的 AML1/ETO 的结合位点。

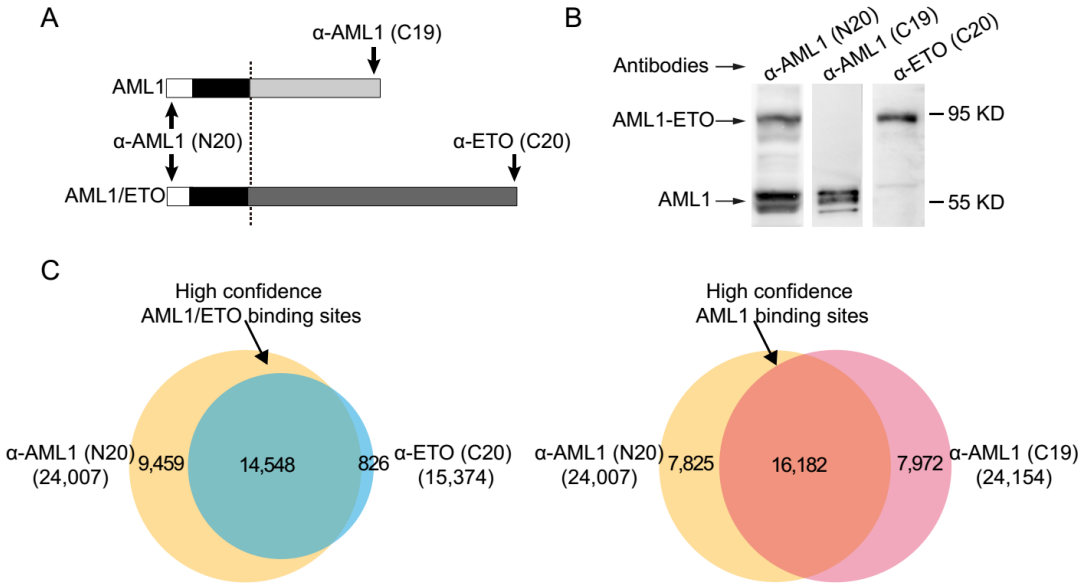


图 3 在 t(8;21)白血病中鉴定出高可信度的 AML1/ETO 和 AML1 结合位点

Figure 3. Identification of high-confidence AML1/ETO and AML1 binding sites in t(8;21) leukemic cells

同时，为了验证我们找到的结合位点，我们下载了多套 AML1/ETO 和 AML1 在 t(8;21)白血病细胞系（Kasumi-1 细胞系和 SKNO-1 细胞系）的 ChIP-seq 数据[21,25,32,33]，并按照相同的流程进行了重新分析。如图 4 所示，大部分鉴定出的高可信度的 AML1/ETO 和 AML1 结合位点 在其他数据中也是 AML1/ETO 或者 AML1 的结合位点，其信号强度的趋势也是相对应的。

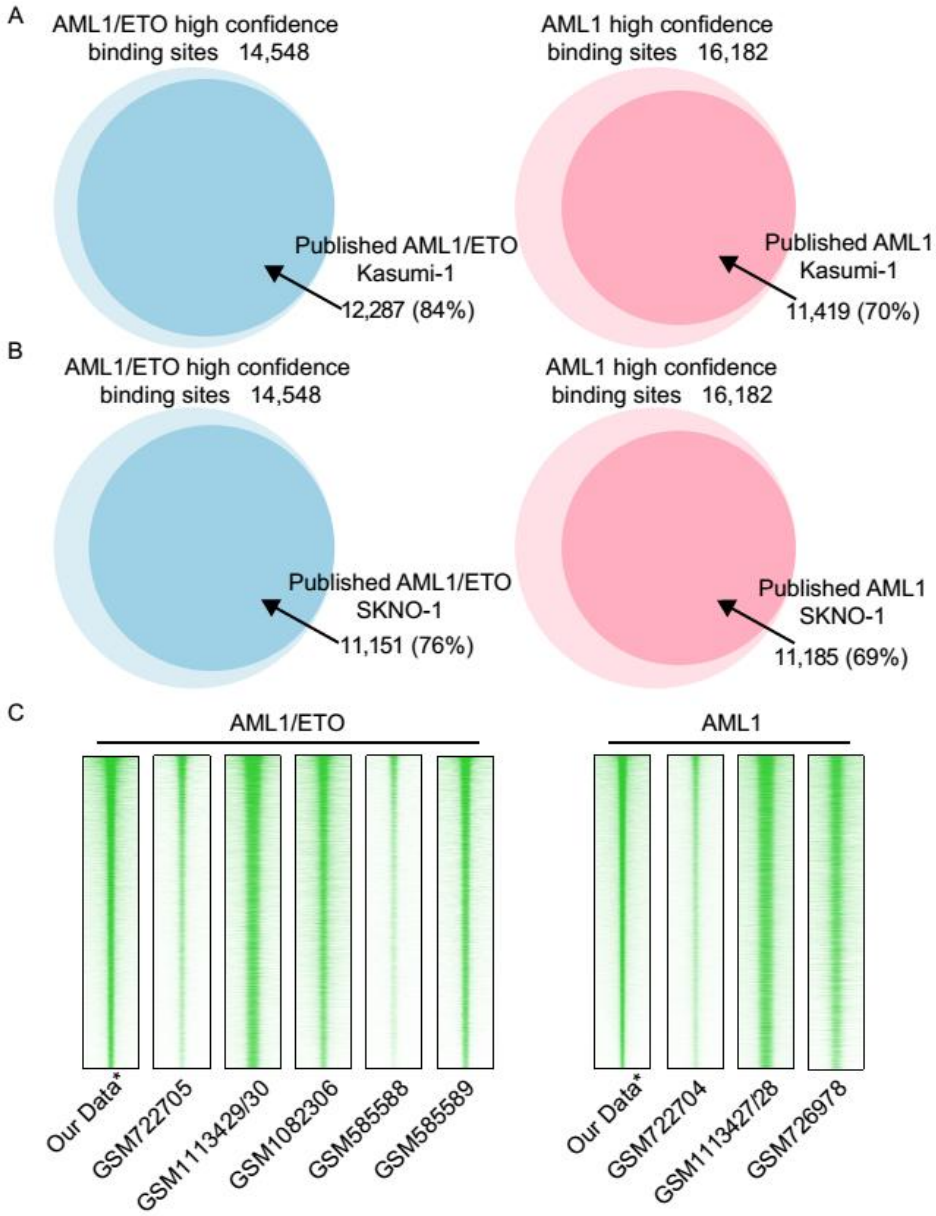


图 4 利用公开发表的 ChIP-seq 数据对鉴定出的高可信度的 AML1/ETO 和 AML1 结

合位点进行验证

Figure 4. The validation of the high confidence AML1/ETO and AML1 binding sites using published ChIP-seq data

3.2 AML1/ETO 与 AML1 的结合位点高度重合，且可以重新分布 AML1 在基因组上的结合

在鉴定了高可信度的 AML1/ETO 和 AML1 的结合位点之后，我们比较了他们在基因组上的结合位置，发现 78%的 AML1/ETO 的结合位点和 70%的 AML1 的结合位点重合（图 5）。这一结果基本上与之前的结果 [25,32,33] 一致，但是由于我们采用了三种抗体鉴定出高可信度的结合位点，我们的结果显示出了更高重合比例。这些结果都提示我们 AML1/ETO 和 AML1 很有可能位于基因组上相同的位置。

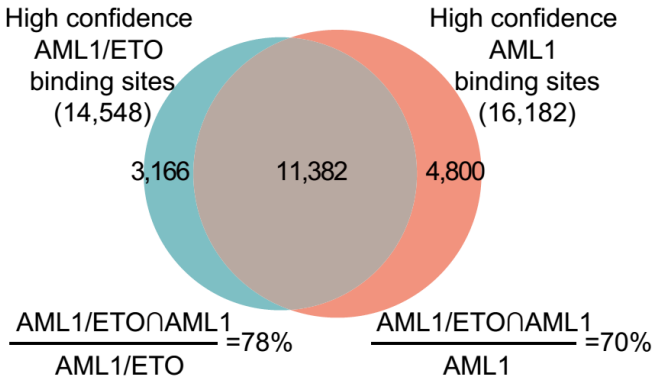


图 5 AML1/ETO 和 AML1 的结合位点高度重合

Figure 5. The binding sites of AML1/ETO and AML1 were highly overlapped

另外，我们比较了野生型的 AML1 蛋白在 AML1/ETO 阳性和阴性细胞中的基因组分布情况。如图 6 所示，野生型的 AML1 在 AML1/ETO 阴性的细胞中的结合位点相较于在 AML1/ETO 阳性的细胞中结合位点更少的结合在启动子区域，而更多的结合在基因间区域。而且野生型 AML1 在 AML1/ETO 敲出前后的分布也有类似的结果。这些都显示，AML1/ETO 可以重新分布 AML1 在基因组上的结合。

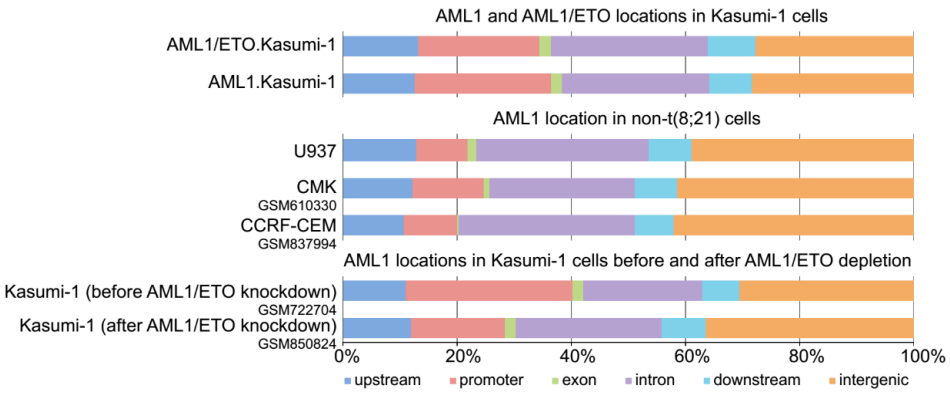


图 6 野生型 AML1 蛋白的结合位点在 AML1/ETO 阳性细胞和阴性细胞中的分布

Figure 6. The genomic distribution of wild-type AML1 in AML1/ETO-positive and –negative cells.

3.3 AML1/ETO 和 AML1 倾向结合临近但是不同的 motif 序列

之后，我们进一步对 AML1/ETO 和 AML1 的重合区域进行研究。由于 ChIP-seq 的最高峰（summit）被认为最可能是蛋白质在 DNA 上的结合位点 [52]（如图 7A），所以我们计算了重叠区域上 AML1/ETO 和 AML1 的 summit 之间的距离。如图 7B 所示，AML1/ETO 和 AML1 的最高峰在同一位置的比例并不高，而大多数情况是存在一定的距离，62%的 peak 中的 summit 的距离在 11-100bp 之间。这提示我们，在重叠的区域 AML1/ETO 和 AML1 很有可能结合在临近的位置。

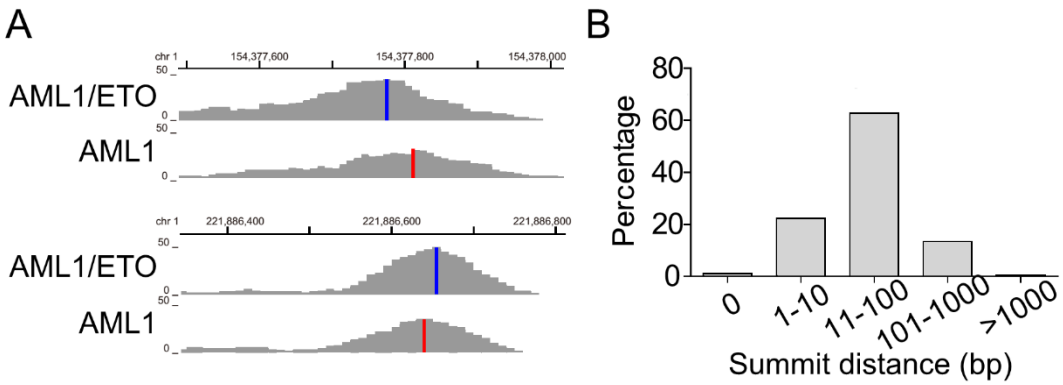
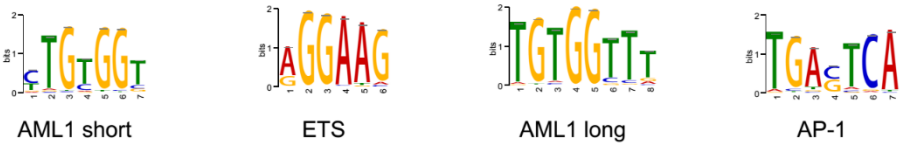


图 7 在重叠的区域 AML1/ETO 和 AML1 的最高峰存在一定距离

Figure 7. The certain distances between the summits of AML1/ETO and AML1 binding existed on the overlap peaks.

我们利用软件 AMD[39] 对于重叠区域中 AML1/ETO summit 周围和 AML1 summit 周围的区域进行了 motif 分析，发现有两种相似但不同的 AML1 的 motif 均显著性的富集了出来（如图 8A），即一个短的 AML1 motif 5'-TG(T/C)GGT-3'和一个长的 AML1 motif 5'-TGTGGTTT-3'。这两个 motif 虽然包含相似的核心序列（TGTGGT）但却有着不一向的侧翼序列，这很有可能是由 AML1/ETO 和 AML1 不同的结构所导致的。另外，AP-1 和 ETS 的 motif 也在被富集了出来（如图 8A）。同时，我们又采用了另外两种 motif 分析软件（HOMER[40] 和 MEME-ChIP[37]）进行了同样的分析，均验证了我们的结果。

A



B

Motif	Best match	AMD				HOMER				MEME-ChIP			
		AML1/ETO		AML1		AML1/ETO		AML1		AML1/ETO		AML1	
		Rank	MAP score	Rank	MAP score	Rank	*P*-value	Rank	*P*-value	Rank	*E*-value	Rank	*E*-value
	AML1	1	17.74	1	17.78	1*	1.0E-2795	2*	1.0E-2956	2*	3.3E-176	2*	1.8E-270
	ETS	2	16.41	2	16.49	2	1.0E-2587	1	1.0E-3115	1	8.3E-661	1	1.4E-788
	AML1	3	15.27	3	14.75	1*	1.0E-2272	2*	1.0E-2289	2*	2.1E-532	2*	9.7E-573
	AP-1	4	14.45	4	14.49	3	1.0E-440	3	1.0E-480	5	9.1E-45	6	1.4E-43

图 8 AML1/ETO 和 AML1 重叠区域的 motif 分析（注：B 图中的*代表 motif 在同一类中）

Figure 8. The motif analysis of AML1/ETO and AML1 overlap regions. * indicates motifs in the same cluster

进一步，我们分析了长短 AML1 motif 在重叠区域中的比例以及在 AML1/ETO 和 AML1 summit 周围的位置分布。如图 9A 所示，大部分（58.73%）AML1/ETO 和 AML1 重叠的区域同时含有这两种 motif。而短的 AML1 motif 更倾向于结合在 AML1/ETO 的周围，而长的 AML1 的 motif 更倾向于结合在 AML1 的周围（如图 9B）。这一结果也被 AML1/ETO 和 AML1 单独区域中所扫到的 motif 所支持（如图 10）。

综合以上结果，我们认为 AML1/ETO 和 AML1 结合在临近的相似的但是不同的 motif 上。

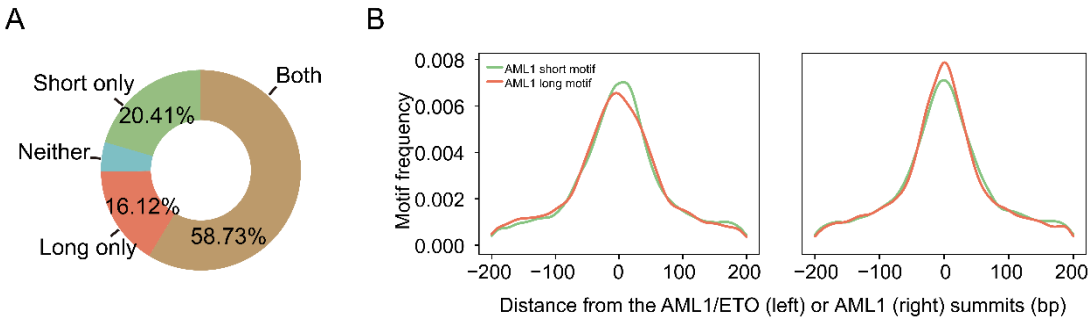


图 9 在重叠区域长短 AML1 motif 所占的比例以及 AML1/ETO 和 AML1 最高峰周围的位置分布

Figure 9. The percentage of the AML1/ETO and AML1 overlapped regions that contain the short and/or long AML1 motifs and the position distribution of the short and long AML1 motifs relative to the AML1/ETO and AML1 peak summits on the overlap regions

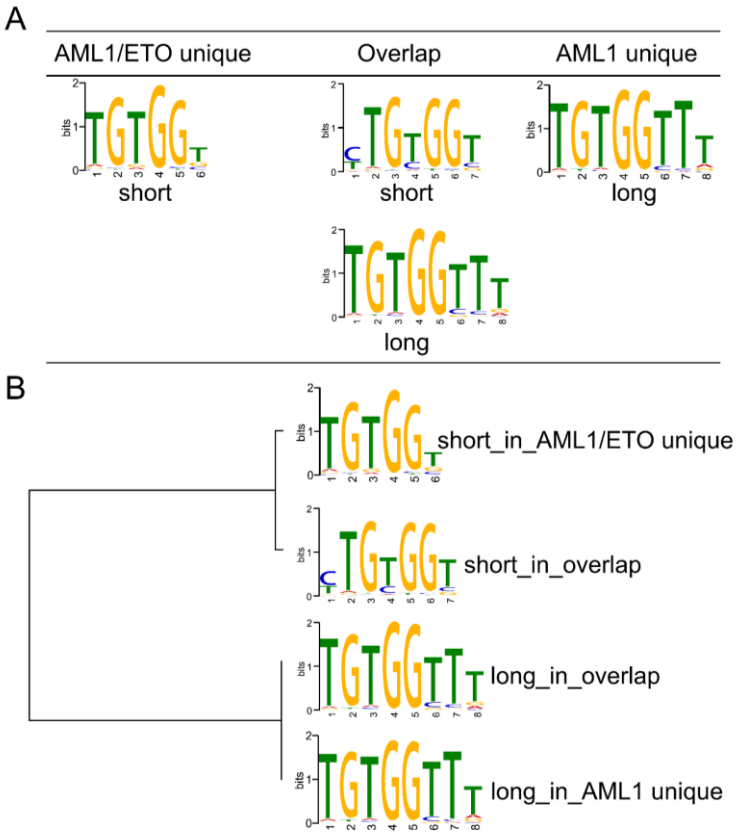


图 10 AML1/ETO 和 AML1 重叠区域与单独区域富集的 AML1 motif 比较

Figure 10. The comparison of AML1 motifs enriched in AML1/ETO unique, AML1 unique and overlap regions

3.4 AML1 存在于 AML1/ETO 形成的复合体中

由于 AML1/ETO 被报道可以与其他因子（如 E2A、HEB、LMO2 等）形成一个稳定的复合体 [12]，因此我们进一步研究是否野生型的 AML1 蛋白也存在于这一 AML1/ETO 形成的复合体中。利用这一报道所公开的 ChIP-seq 数据，我们可以看到我们的 AML1 和 AML1/ETO 的 ChIP-seq 信号与公共数据 [12] 中的 E2A、HEB 和 LMO2 的信号趋势一致（如图 11）且具有相关性（如图 12，对于 AML1 和 AML1/ETO、E2A、HEB 和 LMO 的相关性，R 值分别为 0.663、0.492、0.439 和 0.475，且 p-value 均小于 2.2E-16）。这些结果都提示我们野生型的 AML1 存在于 AML1/ETO 形成的复合体中。

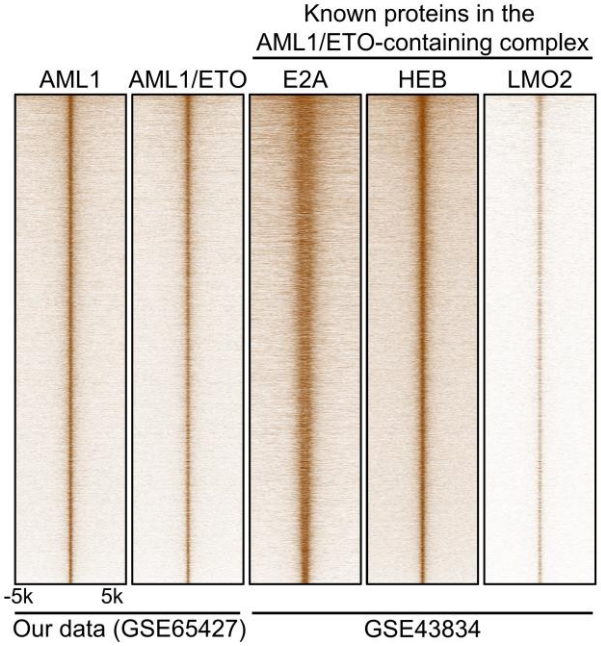


图 11 热图展示 AML1，AML1/ETO，E2A，HEB 和 LMO2 在 AML1/ETO summit 周围（GSE43834）相似的信号分布

Figure 11. Heatmaps showing the similar distribution of binding signals for AML1, AML1/ETO, E2A, HEB and LMO2 on 10bk regions centered on the AML1/ETO

(GSE43834) peak summits

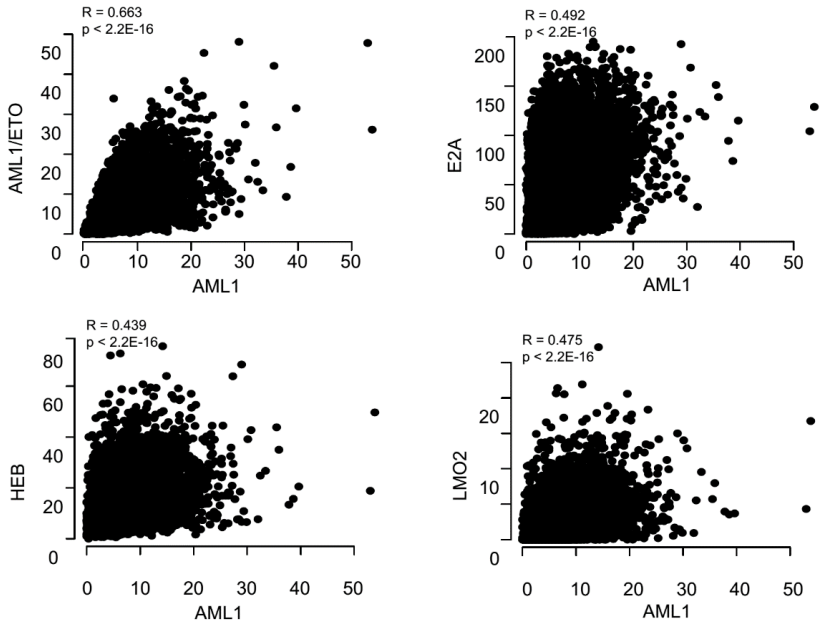


图 12 AML1 与 AML1/ETO、E2A、HEB 和 LMO2 信号强度的相关性分析

Figure 12. Correlation analysis of AML1 binding with AML1/ETO, E2A, HEB and LMO2 binding

3.5 AML1/ETO 与 AML1 之间的相对结合强度与 AML1/ETO 起转录激活或抑制有关

既然 AML1/ETO 与 AML1 可以形成复合体，那么野生型的 AML1 是否参与到了 AML1/ETO 对下游基因的转录调控作用之中了呢。我们首先提取了三套公共数据库[12,32,45]中的 AML1/ETO 的调控基因（详见“材料与方法”）。基于基因集合分析，我们发现 AML1/ETO 下调的基因与 AML1/ETO-AML1 复合体的靶点之间的关联性是显著的（如表 4），这与 AML1/ETO 可以一直 AML1 的靶基因的观点是一致的。然而，我们还发现 AML1/ETO 上调的基因与 AML1/ETO-AML1 符合体的靶点之间也显著性关联，尽管其关联性可能弱于下调的基因。故，我们认为 AML1/ETO-AML1 符合同时参与 AML1/ETO 的上调和下调的转录调控过程。

表 4 AML1/ETO-AML1 复合体的潜在靶点与 AML1/ETO 调节基因集合相关性的富集分析

Table 4. Enrichment analysis of potential targets bound by the AML1/ETO-AML1 complex with gene sets associated with AML1/ETO regulation

Gene Set	Number in the gene set	Potential targets bound by the AML1/ETO-AML1complex			
		Number in the gene set	Fold enrichment	Z-score	P-value
Kasumi-1_AML1/ETO_down	407	232	2.5	14.48	2.87E-35
Kasumi-1_AML1/ETO_up	359	166	2.03	9.32	4.78E-17
SKNO-1_AML1/ETO_down	660	255	2.61	15.92	1.05E-41
SKNO-1_AML1/ETO_up	700	200	1.93	9.45	6.33E-18
AML-M2_ AML1/ETO_down	544	267	2	11.59	1.02E-25
AML-M2_AML1/ETO_up	416	196	1.92	9.33	1.71E-17

由于 AML1/ETO 和 AML1 在重叠区域有不同信号强度的结合（如图 13C），我们采用了 GSEA 分析去评估是否 AML1/ETO 与 AML1 结合强度与 AML1/ETO 的转录调控有关。如图 13A 和 B 所示，AML1/ETO 抑制的基因倾向于有更多 AML1/ETO 的结合。这一组中包含了许多已被报道的 AML1/ETO 主要抑制的基因如 OGG1（图 13C）。另外，我们的结果还显示，AML1/ETO 只是削弱了 AML1 的结合但并没有完全取代，这一观点也被 AML1/ETO 敲除后 AML1 的结合上升的观察所支持（图 14）。而 AML1/ETO 激活的基因则有着更多 AML1 的结合（图 13A 和 B），而 AML1/ETO

的敲出并没有影响 AML1 的结合（图 14）。这可能提示，在激活的作用时，AML1/ETO 起到一个平台作用供其他调节因子结合。

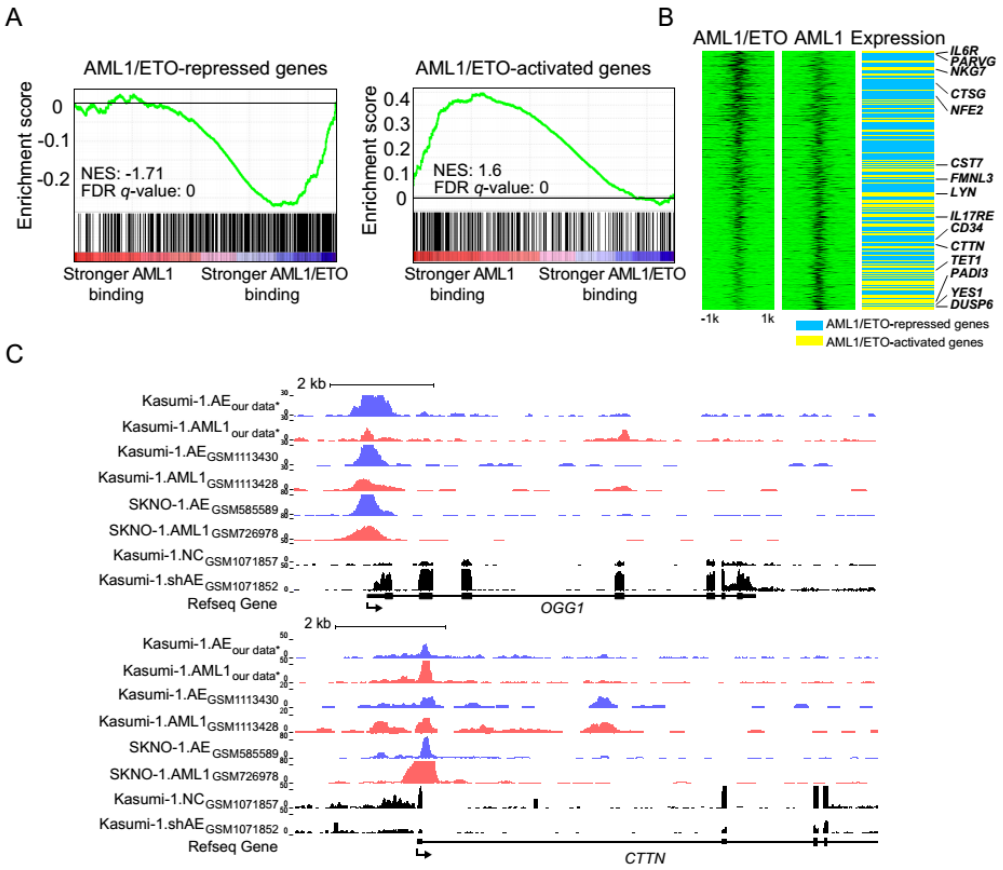


图 13 AML1/ETO 抑制基因 AML1/ETO 的结合信号更强，AML1/ETO 激活基因 AML1 的结合信号更强

Figure 13. AML1/ETO-repressed genes show more AML1/ETO binding and AML1/ETO-activated genes show more AML1 binding

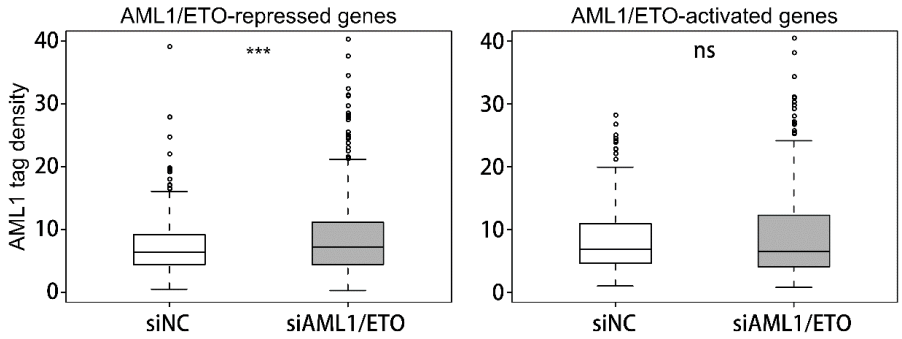


图 14 比较在 AML1/ETO 抑制（左）和激活（右）的基因上 AML1 在 AML1/ETO 敲出前后的结合信号

Figure 14. Comparison of the AML1 binding signals before and after AML1/ETO knockdown on AML1/ETO-repressed genes (the left panel) and AML1/ETO-activated genes (the right panel)

3.6 AP-1 参与 AML1/ETO 的转录激活作用

之后我们进一步研究 AML1/ETO 的激活作用是否有其他转录调节因子的参与。于是，我们采用了两个独立的网络工具，ENCODE ChIP-seq Significance Tool[14] 和 Cscan[15]（详见“材料与方法”）。如表 5 和 6 显示，多个 AP-1 家族的成员富集在 AML1/ETO 激活基因的启动子区域，而在抑制基因的启动子区域却没有发现 AP-1 家族的蛋白。由此，我们预测 AP-1 是 AML1/ETO 转录激活作用的共调节因子。

表 5 利用 ENCODE ChIP-seq Significance Tool 预测 AML1/ETO 激活和抑制基因启动子区域的转录调节因子

Table 5. Transcriptional regulators enriched in the promoter of AML1/ETO-activated and –repressed genes with the ENCODE ChIP-seq Significance Tool

Factor	Total genes with factor	Observed genes (162 in total)	Q-value	Factor rank
AML1/ETO-activated gene				
Ezh2	4510	73	6.04E-09	1
CfosTam112h	6743	89	4.19E-07	2
CfosEtoh01	5461	74	6.37E-06	3
Cfos	7279	89	1.25E-05	4
CfosTam14h	6107	78	1.56E-05	5
Gata2	7233	88	1.56E-05	5
CfosTam	5903	73	1.60E-04	7
Pol2	16037	147	5.26E-04	8

Cjun	5887	69	1.85E-03	9
Cebpb	12640	122	3.23E-03	10
AML1/ETO-repressed gene				
Ebf1	7913	153	4.63E-20	1
Pu1	9665	166	1.37E-17	2
Cmyc	12179	183	1.07E-14	3
Max	14735	202	1.07E-14	3
Pol2	16037	204	4.29E-10	5
Bcl11a	2425	61	6.41E-10	6
Runx3	10961	160	6.04E-09	7
Tcf3	5863	104	1.28E-08	8
Bhlhe40	9666	146	1.36E-08	9
P300	12684	172	1.85E-07	10

表 6 利用 Cscan 预测 AML1/ETO 激活和抑制基因启动子区域的转录调节因子

Table 6. Transcriptional regulators enriched in the promoter of AML1/ETO-activated and –repressed genes with Cscan

Factor	**HYP_BENJ**	**FG_HITS/FG_SIZE**	**Factor rank**
AML1/ETO-activated genes			
FOSL2	4.17E-21	116/166	1
c-Fos	8.42E-20	118/166	2
GATA-2	1.05E-17	94/166	3
c-Jun	9.11E-18	97/166	4
GATA3	4.42E-17	108/166	5
TCF12	5.24E-16	105/166	5
BATF	2.48E-14	95/166	7

IKZF1	2.67E-14	93/166	8
STAT3	2.54E-14	104/166	9
FOSL1	3.09E-14	66/166	10
AML1/ETO-repressed genes			
PU.1	5.40E-22	200/229	1
Max	4.55E-22	161/229	2
EBF1	6.14E-15	146/229	3
PU.1	6.38E-14	157/229	3
PU.1	2.54E-13	140/229	5
c-Myc	6.43E-13	111/229	6
PU.1	1.39E-12	158/229	7
ZBTB7A	1.55E-12	147/229	8
PAX5	1.28E-11	111/229	9
COREST	1.28E-10	209/229	10

4、 结论

（1） 野生型的 AML1 和 AML1 在基因组上结合在相同染色质区域，形成复合体，并结合在临近但是不同的 motif 上；

（2） AML1/ETO 和 AML1 之间信号强度的梯度决定了 AML1/ETO 转录调控的方向；

（3） 生物信息学预测 AP-1 参与了 AML1/ETO 的转录激活作用。

第二部分 整合大规模病人多维数据系统研究急性髓系白血病的预后标志

1、 绪论

癌症并不仅仅是一种单一的疾病而是许多种疾病的总称[ref]。癌症的分类具有相当长的历史，其对于癌症的诊断、治疗都具有相当重要的作用。若人们可以根据分类，判断出癌症病人的预后情况，既可以根据该情况选择不同的治疗方案，提高病人的生存时间。

癌症分类的方法早期局限于形态学，但在人们开发出检测全基因组基因表达情况的 microarray 技术 [5,6] 之后，Golub 等人于 1999 年在 Science 上发表文章报道，在不用任何先验知识的情况下，仅仅用基因表达量就可以将急性髓系白血病（AML）和急性淋巴细胞白血病（ALL）分开，该研究开启了用分子标志对癌症进行分类的先河 [53]。同时，这篇文章也提出，这样的分类不需要全基因组全部基因的表达量，而只需要有限个基因的表达量，即 Gene Signature，即可完成。更严格地说，所谓 Gene Signature 就是指与疾病诊断、预后和治疗反应预测特异性相关的一个或者多个基因的集合。

急性髓系白血病（AML）作为癌症的一种，也是由一组异质性的疾病组成。传统的 AML 分类方法是 1976 年 FAB（French-American-British）组织根据形态学提出的 [54]，之后一直在临床上有广泛的应用，但是 FAB 分型并不能对病人的预后特征给出有效的分类。于是，在 2000 年左右，英国的医学研究协会（medical research council, MRC）、美国的西南肿瘤协作组（Southwest Oncology Group，SWOG）/东部肿瘤协作组（Eastern Cooperative Oncology Group，ECOG）以及美国的癌症与白血病协作组（Cancer and Leukemia Group B）先后发表文章 [55-57]，在上千例病人样本中证明 AML 病人可以根据细胞遗传学特征分为三组——低危组（favourable）、中危组（intermediate）和高危组（unfavourable）。随后，这一结果被多篇文章证实，并收录于 2008 年世界卫生组织（WHO）发表的造血与淋巴组织肿瘤分类中。至此，根据细胞遗传学特征

对白血病进行分类的方法被广泛应用在临床诊断和治疗上，尤其是后期治疗方案的选择上起到了重要的作用。然而，由于 AML 病人中有大约一半的病人核型正常而被归类在中危组中，这些病人的预后需要新的预后指标进行分型。目前，越来越多的文章报道基因突变等分子特征可以成为临床预后指标。例如，CEBPA 的点突变（尤其是两个等位基因同时突变）被报道是预后好的标志[58]、NPM1 的点突变被报道是预后好的标志[59]、DNMT3A 的点突变被报道是预后差的标志[60,61]、FLT3 的突变（串联重复突变和点突变）被报道是预后差的标志[62]等等。

另外利用其它分子层面的数据也可以对 AML 病人进行分类，鉴定出预后标志用来帮助临床对 AML 病人进行诊断、治疗和预后分析也是非常重要的研究方向。2004 年， 新英格兰杂志（NEJM）同时发表两篇文章[63,64]报道用表达谱芯片高通量测定 AML 病人的基因表达情况，并进行了无监督聚类分析和寻找与预后相关的 Gene Signature 研究，将 AML 分成 16 个类别并鉴定出了含有 133 个基因的 Gene Signature。之后，利用不同方法含有不同数量基因的 Gene Signature 被人们鉴定出来。如含有 86 个芯片探针的 Gene Signature[65]，与 FLT3 相关的 Gene Signature[66]，与年龄相关的 Gene Signature[67]，与白血病干细胞相关的 Gene Signature[68]等。另外，人们还证明出 AML 病人的小 RNA 表达量[69]、DNA 甲基化水平[70]、长链非编码 RNA 的表达量[71]也可以对病人进行分类和寻找预后相关的生物标识。

该研究试图整合基因表达量、小 RNA 表达量、DNA 甲基化水平这三维数据，在不同分子特征的 AML 病人中进行分析，找到急性髓系白血病发病相关的预后指标。

2、 材料与方法

2.1 数据的收集与整理

TCGA 计划（癌症基因组地图集计划，The Cancer Genome Atlas）是目前非常著名的大规模病人测序计划，由美国国家癌症研究所和国家人类基因组研究所联合进行，该计划旨在通过大规模的基因组、转录组、表观组、小 RNA 等测序从分子层面进一步认识癌症。目前该计划已经公开发布了超过 30 种癌症的高通量数据，其中就包括急性髓系白血病的数据。2013 年，TCGA 的急性髓系白血病数据正式在新英格兰杂志上发表[72]，共包括 200 例成人急性髓系白血病病人，对其中的 50 例进行了全基因组测序，另外 150 例进行了外显子捕获测序， 200 例进行了 SNP 芯片检测，197 例进行了表达谱芯片检测，179 例进行了 RNA-seq 测序，192 例进行了 DNA 甲基化测序，187 例进行了 microRNA-seq 测序。同时，TCGA 组织还会对这 200 例病人进行随访，提供临床相关数据。

我们希望整合 AML 病人的基因表达、小 RNA 表达和 DNA 甲基化这三维数据。在 200 例病人中，有 175 例病人同时含有这三种数据，我们就使用这 175 例病人样本的数据进行分析。

2.1.1 TCGA AML 病人的基因与小 RNA 数据的收集与整理

TCGA AML 病人的基因表达量数据使用的是 RNA-seq 的 RPKM 值，下载链接为 https://tcga-data.nci.nih.gov/docs/publications/laml_2012/laml.rnaseq.179_v1.0_gaf2.0_rpkm_matrix.txt.tcgaID.txt.gz。TCGA AML 病人的小 RNA 表达量数据使用的是 microRNA-seq 的 RPM 值，下载链接为 https://tcga-data.nci.nih.gov/docs/publications/laml_2012/laml.mirnaseq.rpm.expn_matrix_mimat_norm_passed_TCGA_nodead.txt.gz。另外，为了便于处理和标准化，我们对数据进行了以 2 为底的对数转化和平均值中心缩放处理。

2.1.2 TCGA AML 病人的 DNA 甲基化数据的收集与整理

TCGA 使用了 Illumina HM27K 和 HM450K 两款芯片测量 AML 病人 DNA 甲基化水平，由于 HM450K 监测更多的 CpG 位点，我们选用 HM450K 的数据进行分析。

每个 CpG 的位点使用 β-value 进行定量，β-value 是介于 0-1 之间的一个比值，β-value 为 0 代表该位点在所有细胞中均没有甲基化，β-value 为 1 代表改位点在所有细胞中全部甲基化，下载链接为：https://tcga-data.nci.nih.gov/docs/publications/laml_2012/LAML.HumanMethylation450.Level_3.tgz。另外，为了方便处理，我们对于同一个基因上不同的 CpG 位点的 β-value 值取平均值，来代表这个基因整体的甲基化水平。

2.2 生存分析

由于基因表达量、小 RNA 表达量、DNA 甲基化水平均是连续性变量，故采用 Cox 回归进行生存分析。当利用中位数将病人分成两组后，进行 log-rank 生存分析。生存情况的展示使用 Kapplan-Meier 生存曲线。以上分析和展示均使用 R 语言中的 survival 软件包完成。P value 小于 0.05 被认为是具有显著性差异。ROC 曲线下面积（AUC）分析采用的是 R 语言的 survivalROC 包完成。AUC 值越高代表模型越准确，AUC 值若为 0.5 代表模型与随机无差别，AUC 值若为 1 代表模型完美。计算细胞遗传学分类的 AUC 值时，高风险、中风险和低风险的分数被设置为 3,2,1。

2.3 训练集和检验集

为了使得我们的结果能够得到验证，一般在做数据分析的时候，均会将数据分成训练集（training set）和检验集（validation set）。在训练集中训练数据，建立模型，在检验集中进行验证。在这次研究中，我随机挑选 175 例病人样本中的 60%的数据（即 105 例病人）作为训练集，剩下的 40%的数据（70 例）作为检验集。随机挑选的步骤使用 R 语言中的 sample 函数完成。

2.4 主成分分析

我们利用主成分分析对 25 个分子特征（10 个 mRNA、5 个小 RNA 和 10 个 DNA 甲基化水平）进行整合和打分，第一主成分作为这些分子特征的权重参数，最后根据 score = $\sum value_i * weight_i$ 进行计算。主成分分析利用 R 语言中的 prcomp 函数完成。

3、 结果

3.1 病人样本情况

我们收集整理了 175 例 AML 病人样本的基因表达量、小 RNA 表达量和 DNA 甲基化水平的数据。并将 175 例 AML 病人样本分出训练集（105 例病人）和验证集（70 例病人）两部分（如图 15）。其具体的临床信息、FAB 分型、细胞生物学分型和重要的分子特征如表 7。

表 7 TCGA AML 病人样本的基本情况

Table 7. Summary of TCGA AML samples

	Total samples	**Training samples**	**Validation samples**
Sample number	175	105	70
Age, median (IQR)	58 (45-67)	56 (45-65)	60 (46-69)
Gender			
Male	91	57	34
Female	84	48	36
Cytogenetic risk group			
Favorable	32	22	10
Intermediate	98	59	39
Unfavorable	42	21	21
Missing data	3	3	0
AML FAB subtype			
M0	16	9	7
M1	43	24	19
M2	39	26	13
M3	15	12	3
M4	35	18	17
M5	20	12	8

M6	2	2	0
M7	3	0	3
Other subtype	2	2	0
Normal karyotype	78	45	33
DNMT3A mutation	41	23	18
FLT3 mutation	48	26	22
NPM1 mutation	46	24	22

IQR, interquartile range.

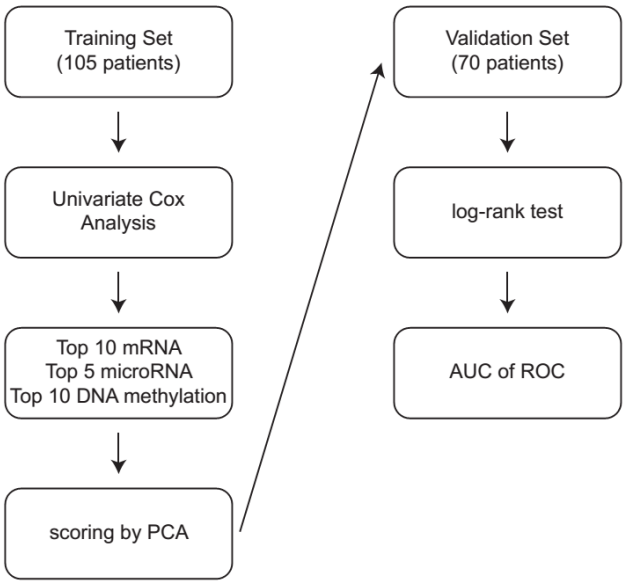


图 15 生存分析流程图

Figure 15. The schematic flow chart of survival analysis

3.2 筛选预后标志的基因、小 RNA 和 DNA 甲基化

针对训练集中的 105 例 AML 病人，我们对 19990 个基因、865 个小 RNA 和 19013 个基因的 DNA 甲基化水平进行了单因素的 Cox 生存回归分析。为了使得找到的分子标志能在不同分子特征的病人中均对预后有统计学意义。我们又分别在正常核型、DNMT3A 突变和没突变、FLT3 突变和没突变、NPM1 突变和没突变这七种情况做了相同的分析。加上所有病人这种情况，我们筛选在这八种分子特征的病人集合中，

至少 5 种具有显著性差异的特征基因、小 RNA 和 DNA 甲基化。结果找到了与 AML 预后相关的 829 个基因（其中 680 个高表达与预后差相关、149 个高表达与预后好相关），24 个小 RNA（其中 21 个高表达与预后差相关、3 个高表达与预后好相关）和 92 个基因的甲基化水平（其中 16 个基因的高甲基化水平与预后差相关、76 个基因的低价计划水平与预后差相关）。表 10 中列出了显著性排名前 10 的基因、前 5 的小 RNA 和前 10 的基因的甲基化。

表 10 与预后最显著性相关的基因、小 RNA 和 DNA 甲基化

Table 10. Most significant genes, microRNA and DNA methylation associated with prognosis

Rank	Moclular features	HR (95%CI)	P value	Type
Top 10 gene expressions				
1	SLC17A3	12.94 (5.15-32.54)	5.27E-08	Risk
2	MUC13	1.83E+11 (8.90E+06-3.77E+15)	3.09E-07	Risk
3	CLEC11A	0.78 (0.71-0.86)	5.91E-07	Pro
4	MPO	0.83 (0.77-0.89)	6.30E-07	Pro
5	RFX6	7.55E+39 (5.26E+23-1.08E+56)	1.32E-06	Risk
6	MYOC	17.70 (5.52-56.75)	1.35E-06	Risk
7	AMBN	523.08 (40.53-6750.40)	1.61E-06	Risk
8	SERPINI1	2.49 (1.71-3.64)	2.32E-06	Risk
9	KIAA0125	1.39 (1.21-1.59)	2.67E-06	Risk
10	PTPRA	21.77 (5.93-79.87)	3.40E-06	Risk
Top 5 microRNA expressions				
1	hsa.mir.106a	1.41 (1.20-1.65)	1.80E-05	Risk
2	hsa.mir.20b	1.26 (1.13-1.42)	4.65E-05	Risk
3	hsa.mir.363	1.26 (1.13-1.42)	7.06E-05	Risk

4	hsa.mir.532	1.57 (1.26-1.97)	7.74E-05	Risk
5	hsa.mir.20b	1.25 (1.12-1.39)	9.05E-05	Risk
Top 10 DNA methylation				
1	E2F7	0 (0-0)	1.51E-07	Pro
2	CALR	790.97 (50.39-12416.56)	2.03E-06	Risk
3	MFHAS1	0 (0-0)	8.84E-06	Pro
4	PPM1B	0 (0-0)	6.91E-05	Pro
5	EPSTI1	0 (0-0)	8.61E-05	Pro
6	C17orf105	176.20 (13.00-2387.42)	1.01E-04	Risk
7	SPINK2	0 (0-0.06)	1.20E-04	Pro
8	ABP1	261.08 (15.25-4470.05)	1.23E-04	Risk
9	LTB	0 (0-0.02)	1.38E-04	Pro
10	ARID2	0 (0-0)	2.87E-04	Pro

CI, confidence interval; HR, hazard ratio; Pro, prognostic factors; Risk, risk factors.

3.3 整合三维数据的预后标志

通过单因素 Cox 分析找到有预后相关性的基因、小 RNA 和 DNA 甲基化之后，我们将上述最显著的 10 个基因、5 个小 RNA 的表达量、10 个基因的 DNA 甲基化水平整合，利用主成分分析得出权重，进行打分（详见材料与方法 2.4），然后再在检验集中进行验证。如图 16 可以看出，这一整合的预后标志可以很好的对于检验集中的 AML 病人的预后情况进行预测。为了评价整合的预后指标，我们计算了 ROC 曲线下面积，AUC 达到了 0.753（如图 17A），其对预后预测的效果好于基于细胞遗传学的分类（如图 17B）。

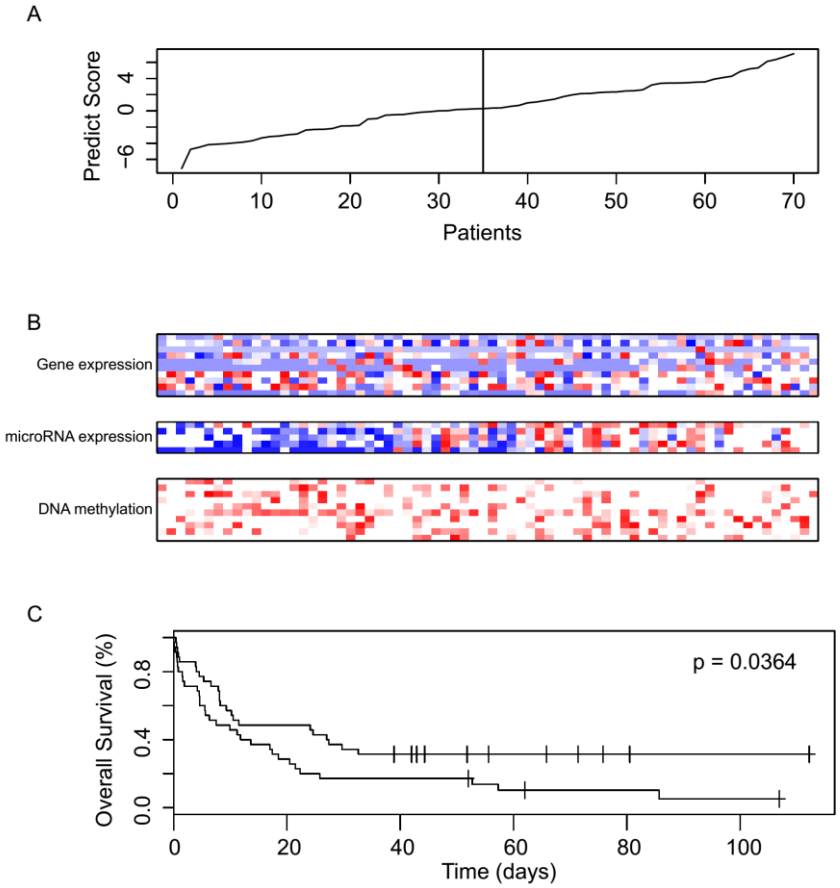


图 16 在训练集中（A 和 B）整合基因、小 RNA 和 DNA 甲基化数据鉴定预后标志并在验证集中（C）进行验证

Figure 16. Integration of gene, microRNA and DNA methylation data to identify the biomarker in training set (A and B) and verification in the validataion set (C)

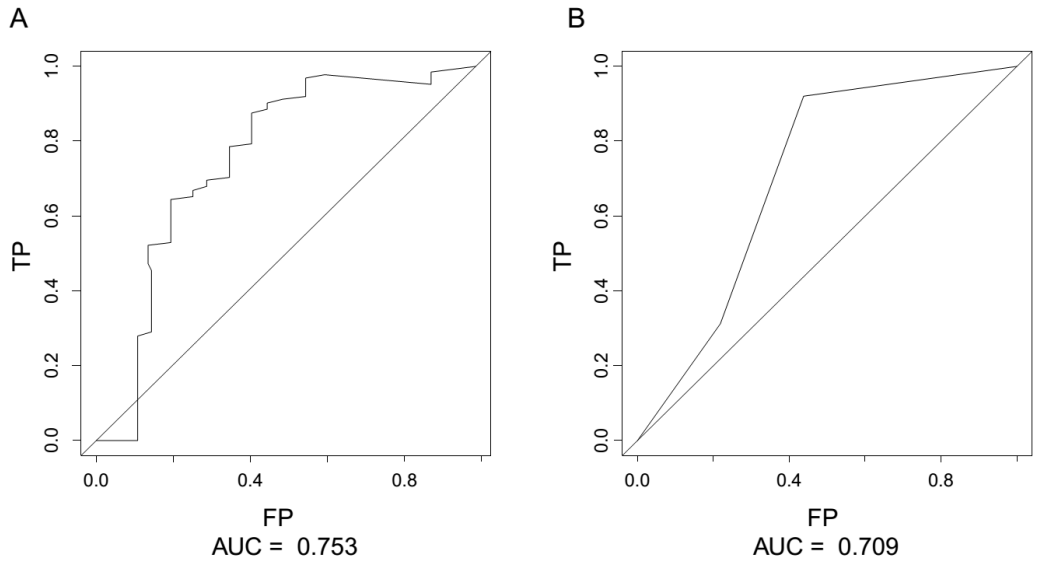


图 17 ROC 曲线下面积（A）为该研究整合的预后指标（B）为细胞遗传学分类

Figure 17. Area under curve (AUC) of receiver-operating characteristic (ROC) curves for this study prognosis factor (A) and cytogentics classification (B)

4、 结论

（1） 鉴定在各种分子特征的 AML 病人中均具有预后意义的基因、小 RNA 和 DNA 甲基化标志；

（2） 整合基因、小 RNA 和 DNA 甲基化三维数据，鉴定出 AML 病人的预后标志。

参考文献

1. Marx V. Biology: The big challenges of big data. *Nature.* Jun 13 2013;498(7453):255-260.
2. Hogeweg P. The roots of bioinformatics in theoretical biology. *PLoS computational biology.* Mar 2011;7(3):e1002021.
3. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. *Journal of theoretical biology.* Mar 1965;8(2):357-366.
4. Ouzounis CA, Valencia A. Early bioinformatics: the birth of a discipline--a personal view. *Bioinformatics.* Nov 22 2003;19(17):2176-2190.
5. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* Oct 20 1995;270(5235):467-470.
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science.* Oct 20 1995;270(5235):484-487.
7. Glenn TC. Field guide to next-generation DNA sequencers. *Molecular ecology resources.* Sep 2011;11(5):759-769.
8. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. *Cell.* Mar 14 2013;152(6):1237-1251.
9. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics.* Oct 2009;10(10):669-680.
10. Bailey T, Krajewski P, Ladunga I, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology.* 2013;9(11):e1003326.
11. Qian M, Jin W, Zhu X, et al. Structurally differentiated cis-elements that interact with PU.1 are functionally distinguishable in acute promyelocytic leukemia. *Journal of hematology & oncology.* 2013;6:25.
12. Sun XJ, Wang Z, Wang L, et al. A stable transcription factor complex nucleated by oligomeric AML1-ETO controls leukaemogenesis. *Nature.* Aug 1 2013;500(7460):93-97.
13. Walz S, Lorenzin F, Morton J, et al. Activation and repression by oncogenic MYC shape tumour-specific gene expression profiles. *Nature.* Jul 24 2014;511(7510):483-487.
14. Auerbach RK, Chen B, Butte AJ. Relating genes to function: identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool. *Bioinformatics.* Aug 1 2013;29(15):1922-1924.
15. Zambelli F, Prazzoli GM, Pesole G, Pavesi G. Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. *Nucleic acids research.* Jul 2012;40(Web Server issue):W510-515.
16. Lam K, Zhang DE. RUNX1 and RUNX1-ETO: roles in hematopoiesis and leukemogenesis. *Frontiers in bioscience.* 2012;17:1120-1139.
17. Helbling D, Mueller BU, Timchenko NA, et al. The leukemic fusion gene AML1-MDS1-EVI1 suppresses CEBPA in acute myeloid leukemia by activation of Calreticulin. *Proceedings of the National Academy of Sciences of the United States of America.* Sep 7 2004;101(36):13312-13317.

18. Guastadisegni MC, Lonoce A, Impera L, et al. CBFA2T2 and C20orf112: two novel fusion partners of RUNX1 in acute myeloid leukemia. *Leukemia.* Aug 2010;24(8):1516-1519.
19. Peterson LF, Zhang DE. The 8;21 translocation in leukemogenesis. *Oncogene.* May 24 2004;23(24):4255-4262.
20. Hatlen MA, Wang L, Nimer SD. AML1-ETO driven acute leukemia: insights into pathogenesis and potential therapeutic approaches. *Frontiers of medicine.* Sep 2012;6(3):248-262.
21. Liu Y, Chen W, Gaudet J, et al. Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity. *Cancer cell.* Jun 2007;11(6):483-497.
22. Liu Y, Cheney MD, Gaudet JJ, et al. The tetramer structure of the Nervy homology two domain, NHR2, is critical for AML1/ETO's activity. *Cancer cell.* Apr 2006;9(4):249-260.
23. Ben-Ami O, Friedman D, Leshkowitz D, et al. Addiction of t(8;21) and inv(16) acute myeloid leukemia to native RUNX1. *Cell reports.* Sep 26 2013;4(6):1131-1143.
24. Goyama S, Schibler J, Cunningham L, et al. Transcription factor RUNX1 promotes survival of acute myeloid leukemia cells. *The Journal of clinical investigation.* Sep 2013;123(9):3876-3888.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* Jul 15 2009;25(14):1754-1760.
26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods.* Apr 2012;9(4):357-359.
27. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* Aug 15 2009;25(16):2078-2079.
28. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology.* Dec 2008;26(12):1351-1359.
29. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome biology.* 2008;9(9):R137.
30. Qin ZS, Yu J, Shen J, et al. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC bioinformatics.* 2010;11:369.
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* Mar 15 2010;26(6):841-842.
32. Martens JH, Mandoli A, Simmer F, et al. ERG and FLI1 binding sites demarcate targets for aberrant epigenetic regulation by AML1-ETO in acute myeloid leukemia. *Blood.* Nov 8 2012;120(19):4038-4048.
33. Ptasinska A, Assi SA, Mannari D, et al. Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia.* Aug 2012;26(8):1829-1841.
34. Pencovich N, Jaschek R, Tanay A, Groner Y. Dynamic combinatorial interactions of RUNX1 and cooperating partners regulates megakaryocytic differentiation in cell line models. *Blood.* Jan 6 2011;117(1):e1-14.
35. Sanda T, Lawton LN, Barrasa MI, et al. Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. *Cancer cell.* Aug 14 2012;22(2):209-221.
36. D'Haeseleer P. What are DNA sequence motifs? *Nature biotechnology.* Apr 2006;24(4):423-425.

37. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics.* Jun 15 2011;27(12):1696-1697.
38. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics.* Jun 15 2011;27(12):1653-1659.
39. Shi J, Yang W, Chen M, Du Y, Zhang J, Wang K. AMD, an automated motif discovery tool using stepwise refinement of gapped consensuses. *PloS one.* 2011;6(9):e24576.
40. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell.* May 28 2010;38(4):576-589.
41. Matys V, Kel-Margoulis OV, Fricke E, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research.* Jan 1 2006;34(Database issue):D108-110.
42. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. *Genome biology.* 2007;8(2):R24.
43. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic acids research.* Jul 2007;35(Web Server issue):W253-258.
44. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics.* Apr 1 2011;27(7):1017-1018.
45. Taskesen E, Bullinger L, Corbacioglu A, et al. Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood.* Feb 24 2011;117(8):2469-2475.
46. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology.* 2013;14(4):R36.
47. Feng J, Meyer CA, Wang Q, Liu JS, Shirley Liu X, Zhang Y. GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics.* Nov 1 2012;28(21):2782-2788.
48. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* Apr 2003;4(2):249-264.
49. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America.* Apr 24 2001;98(9):5116-5121.
50. Wang K, Wang P, Shi J, et al. PML/RARalpha targets promoter regions containing PU.1 consensus and RARE half sites in acute promyelocytic leukemia. *Cancer cell.* Feb 17 2010;17(2):186-197.
51. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America.* Oct 25 2005;102(43):15545-15550.
52. Wang J, Zhuang J, Iyer S, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research.* Sep 2012;22(9):1798-1812.
53. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class

prediction by gene expression monitoring. *Science.* Oct 15 1999;286(5439):531-537.

54. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *British journal of haematology.* Aug 1976;33(4):451-458.
55. Grimwade D, Walker H, Oliver F, et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood.* Oct 1 1998;92(7):2322-2333.
56. Slovak ML, Kopecky KJ, Cassileth PA, et al. Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. *Blood.* Dec 15 2000;96(13):4075-4083.
57. Byrd JC, Mrozek K, Dodge RK, et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood.* Dec 15 2002;100(13):4325-4336.
58. Wouters BJ, Lowenberg B, Erpelinck-Verschueren CA, van Putten WL, Valk PJ, Delwel R. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood.* Mar 26 2009;113(13):3088-3091.
59. Thiede C, Koch S, Creutzig E, et al. Prevalence and prognostic impact of NPM1 mutations in 1485 adult patients with acute myeloid leukemia (AML). *Blood.* May 15 2006;107(10):4011-4020.
60. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *The New England journal of medicine.* Dec 16 2010;363(25):2424-2433.
61. Yan XJ, Xu J, Gu ZH, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nature genetics.* Apr 2011;43(4):309-315.
62. Gale RE, Green C, Allen C, et al. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood.* Mar 1 2008;111(5):2776-2784.
63. Valk PJ, Verhaak RG, Beijen MA, et al. Prognostically useful gene-expression profiles in acute myeloid leukemia. *The New England journal of medicine.* Apr 15 2004;350(16):1617-1628.
64. Bullinger L, Dohner K, Bair E, et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England journal of medicine.* Apr 15 2004;350(16):1605-1616.
65. Metzeler KH, Hummel M, Bloomfield CD, et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood.* Nov 15 2008;112(10):4193-4201.
66. Bullinger L, Dohner K, Kranz R, et al. An FLT3 gene-expression signature predicts clinical outcome in normal karyotype AML. *Blood.* May 1 2008;111(9):4490-4495.
67. de Jonge HJ, de Bont ES, Valk PJ, et al. AML at older age: age-related gene expression profiles reveal a paradoxical down-regulation of p16INK4A mRNA with prognostic significance. *Blood.* Oct 1 2009;114(14):2869-2877.

68. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *Jama.* Dec 22 2010;304(24):2706-2715.
69. Jongen-Lavrencic M, Sun SM, Dijkstra MK, Valk PJ, Lowenberg B. MicroRNA expression profiling in relation to the genetic heterogeneity of acute myeloid leukemia. *Blood.* May 15 2008;111(10):5078-5085.
70. Figueroa ME, Lugthart S, Li Y, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer cell.* Jan 19 2010;17(1):13-27.
71. Garzon R, Volinia S, Papaioannou D, et al. Expression and prognostic impact of lncRNAs in acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America.* Dec 30 2014;111(52):18679-18684.
72. Cancer Genome Atlas Research N. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *The New England journal of medicine.* May 30 2013;368(22):2059-2074.

附录

附录 1 ChIP-seq 流程化分析代码

```
##author: zjuwhw
##time: 2014-12-25
import os,sys
import os.path
import time
def countFilelines(file):
    count=0
    f=open(file,"r")
    for line in f:
        count += 1
    return count
def sra2fastq(srafile):
    if os.path.exists(srafile.replace(".sra",".fastq")) :
        print "The file "+srafile.replace(".sra",".fastq")+" has exist"
    elif os.path.exists(srafile.replace(".sra",".fq")):
        print "The file "+srafile.replace(".sra",".fq")+" has exist"
    else:
        cmd="fastq-dump "+srafile
        os.system(cmd)
def fastq2bam(fastqfile):
    postfix="."+os.path.basename(fastqfile).split(".")[-1]
    ref="/d/database/hg19/bwaindex/hg19.fa"
    if os.path.exists(fastqfile.replace(postfix,".bam")) :
        print "The file "+fastqfile.replace(postfix,".bam")+" has exist"
    else:
        cmd1="bwa aln -t 4 "+ref+" "+fastqfile+" > "+fastqfile.replace(postfix,".sai")
        cmd2="bwa samse "+ref+" "+fastqfile.replace(postfix,".sai")+" "+fastqfile+" >
"+fastqfile.replace(postfix,".sam")
        cmd3="samtools view -bS "+fastqfile.replace(postfix,".sam")+" >
"+fastqfile.replace(postfix,"_unsorted.bam")
        cmd4="samtools sort "+fastqfile.replace(postfix,"_unsorted.bam")+"
"+fastqfile.replace(postfix,"")
        cmd5="samtools index "+fastqfile.replace(postfix,"")+".bam"
        cmd6="rm -f *sam *sai *_unsorted.bam"
        os.system(cmd1)
        os.system(cmd2)
```

```
            os.system(cmd3)
            os.system(cmd4)
            os.system(cmd5)
            os.system(cmd6)

def bam2uniq_nodup_bam(bamfile):
    postfix="."+os.path.basename(bamfile).split(".")[-1]
    if os.path.exists(bamfile.replace(postfix,"_uniq_nodup.bam")):
        print "The file "+bamfile.replace(postfix,"_uniq_nodup.bam")+" has exist"
    else:
        cmd1="samtools view -bq 1 "+bamfile+" > "+bamfile.replace(postfix,"_uniq.bam")
        cmd2="samtools rmdup -s "+bamfile.replace(postfix,"_uniq.bam")+"
"+bamfile.replace(postfix,"_uniq_nodup.bam")
        os.system(cmd1)
        os.system(cmd2)
if __name__ == '__main__':
    if len(sys.argv) < 2:
        print 'No action specified.'
        sys.exit()
    if sys.argv[1].startswith('--'):
        option = sys.argv[1][2:]
        if option=="help":
            print '"\
This program is used for ChIP-seq data preprocess and alignment by bwa software;
usage:
    python chipseq_pipeline.py [srafile or fastqfile or bamfile]
Notes:
    The program "bwa","samtools","sratoolkit" must be installed and avaliable in the PATH.
    The path of reference genome must be set and the bwa index must be bulit.'"
        else :
            print 'Unknown option.'
        sys.exit()
    else:
        for filename in sys.argv[1:]:
            start=time.time()
            postfix="."+os.path.basename(filename).split(".")[-1]
            sra2fastq(filename.replace(postfix,".sra"))
            if os.path.exists(filename.replace(postfix,".fastq")):
                fastq2bam(filename.replace(postfix,".fastq"))
            else:
```

```
        fastq2bam(filename.replace(postfix,".fq"))
bam2uniq_nodup_bam(filename.replace(postfix,".bam"))
end=time.time()
print "The file "+filename+" has done using "+str(end-start)+" seconds"
```

附录 2 ChIP-seq 数据 peak calling 代码

```
#author: zjuwhw
#date: 2014-12-26
USAGE='"peak_calling.py --- peak calling for single-end ChIP-seq bam file using MACS and Hpeak software, and output the bigWig and peak region files
USAGE:
    python %s [--tool="macs","hpeak"] [--gsize=#] [--pvalue=#] [--output=#] [--chromsize=#]sample_bam_file control_bam_file(optional)
Default:
    This pipeline intergated two peak calling software, macs and hpeak. The tool parameter musted be set.
    For macs, the default p-value is 1e-8, and for hpeak, the default p-value is 1e-4
    For macs, the output are bigWig file and _p_s.bed file; for hpeak, the output is hpeak default output, further process should be made by yourself.
    the chromesize file is /c/wanghw/annotation/hg19.chrom.sizes, which could be downloaded from ucsc website.
    the option gsize for macs14, Effective genome size. It can be 1.0e+9 or 1000000000,or shortcuts:'hs' for human (2.7e9), 'mm' for mouse(1.87e9), 'ce' for C. elegans (9e7) and 'dm' for fruitfly (1.2e8), Default:hs
Note:
    This pipeline needs the software macs, hpeak and ucsc's utility wigToBigWig installed in the PATH
    This pipeline was only tested using human data. You must pay more attention dealing with other species' data'"
import os,sys,getopt,time
def macs_xls2psbed(macsxlsfile):
    fxls = open(macsxlsfile)
    fpsbed = open(macsxlsfile.replace("_peaks.xls","_p_s.bed"),"w")
    n = 0
    for line in fxls:
        if line.startswith("chr") and (not line.startswith("chr\t")):
            n = n+1
            linelist = line.rstrip().split("\t")
            print >> fpsbed, "%s\t%d\t%d\t%s\t%d" % (linelist[0], int(linelist[1])-1, int(linelist[2]), "MACS_peak_%s" % n, int(linelist[1])-1+int(linelist[4]) )
    fxls.close()
    fpsbed.close()
def macs_wig2bigWig(name, chromsizefile):
    cmd = "for gzwigfile in %s_MACS_wiggle/treat/*wig.gz;do gunzip -c $gzwigfile| awk 'NR>1{print $0}' >> %s;done" % (name, name + ".wig")
```

```
    os.system(cmd)
    cmd2 = "wigToBigWig -clip %s %s %s" % (name + ".wig", chromsizefile, name + ".bigWig")
    os.system(cmd2)
    cmd3 = "rm -rf %s_MACS_wiggle %s.wig" % (name, name)
    os.system(cmd3)
def macs_nocontrol(samplebamfile, pvalue, name, chromsize, gsize):
    cmd = "macs14 -t " + samplebamfile + " -p " + str(pvalue) + " -n " + name + " -w --space=10 -g "
+ gsize
    os.system(cmd)
    cmd_rm="rm -f *r *negative*xls *peaks.bed *summits.bed *model.r"
    os.system(cmd_rm)
    macs_xls2psbed(name + "_peaks.xls")
    macs_wig2bigWig(name, chromsize)
def macs_havecontrol(samplebamfile, controlbamfile, pvalue, name, chromsize, gsize):
    cmd = "macs14 -t " + samplebamfile + " -c " + controlbamfile + " -p " + str(pvalue) + " -n " +
name + " -w --space=10 -g " + gsize
    os.system(cmd)
    cmd_rm="rm -f *r *negative*xls *peaks.bed *summits.bed *model.r"
    os.system(cmd_rm)
    macs_xls2psbed(name + "_peaks.xls")
    macs_wig2bigWig(name, chromsize)
def bam2hpeakbed(bamfile):
    cmd="bedtools bamtobed -i %s | cut -f1-3,6 > %s" %(bamfile,
os.path.basename(bamfile).replace(".bam",".hpeak.bed"))
    os.system(cmd)
def hpeakout2psbed(name):
    cmd = '''awk '{if($1>=1 && $1<=22){$1="chr"$1}else if($1==23){$1="chrX"}else
if($1==24){$1="chrY"};$2=int($2)-1;$5=int($5)+$2;$4="hpeak_peak_"NR;print
$1"\t"$2"\t"$3"\t"$4"\t"$5}' %s >> %s ''' % (name+".hpeak.out", name+".hpeak_p_s.bed")
    #print cmd
    os.system(cmd)
def hpeak_nocontrol(samplebamfile, pvalue, name):
    bam2hpeakbed(samplebamfile)
    f1=open(name + "_treat.txt","w")
    print >> f1, os.path.basename(samplebamfile).replace(".bam",".hpeak.bed")
    f1.close()
    cmd="perl /c/wanghw/software/HPeak-2.1/HPeak.pl -format BED -t %s_treat.txt -n %s -fmin 100
-fmax 300 -w 25 -s %s" %(name, name, pvalue)
    os.system(cmd)
    os.remove(name + "_treat.txt")
```

```
    os.remove(os.path.basename(samplebamfile).replace(".bam",".hpeak.bed"))
    hpeakout2psbed(name)
def hpeak_havecontrol(samplebamfile, controlbamfile, pvalue, name):
    bam2hpeakbed(samplebamfile)
    bam2hpeakbed(controlbamfile)
    f1=open(name + "_treat.txt","w")
    f2=open(name + "_control.txt","w")
    print >> f1, os.path.basename(samplebamfile).replace(".bam",".hpeak.bed")
    print >> f2, os.path.basename(controlbamfile).replace(".bam",".hpeak.bed")
    f1.close()
    f2.close()
    cmd="perl /c/wanghw/software/HPeak-2.1/HPeak.pl -format BED -t %s_treat.txt
-c %s_control.txt -n %s -fmin 100 -fmax 300 -w 25 -s %s" %(name, name, name, pvalue)
    os.system(cmd)
    os.remove(name + "_treat.txt")
    os.remove(name + "_control.txt")
    os.remove(os.path.basename(samplebamfile).replace(".bam",".hpeak.bed"))
    os.remove(os.path.basename(controlbamfile).replace(".bam",".hpeak.bed"))
    hpeakout2psbed(name+"_treated")
if __name__ == '__main__':
    starttime=time.time()
    if len(sys.argv) < 2:
        print USAGE % sys.argv[0]
        sys.exit(1)
    opts, args = getopt.getopt(sys.argv[1:], "", ["tool=","pvalue=","output=","gsize=", "chromsize="])
    if len(args) > 2:
        print "you can only input 1-2 bam file"
        sys.exit(1)
    if (('--tool','macs') not in opts) and (('--tool','hpeak') not in opts):
        print "the parameter tool must be set, and only the macs or hpeak could be used"
        sys.exit(1)
    # defaults
    def_pval={"macs": "1e-8", "hpeak": "1e-4"}
    nname = os.path.basename(args[0]).replace(".bam","")
    chrom_size = "/c/wanghw/annotation/hg19.chrom.sizes"
    gsize = "hs"
    for o,a in opts:
        if o == '--tool':
            ntype = a
            npvalue = def_pval[a]
```

```
            elif o == '--pvalue':
                npvalue = a
            elif o == '--output':
                nname = a
            elif o == '--chromsize':
                chrom_size = a
            elif o == '--gsize':
                gsize = a
        print r'''#########''', args[0], npvalue, nname, chrom_size, r'''###############'''

        if len(args) == 2 and ntype == "macs":
            macs_havecontrol(args[0], args[1], npvalue, nname, chrom_size, gsize)
        elif len(args) == 1 and ntype == "macs":
            macs_nocontrol(args[0], npvalue, nname, chrom_size, gsize)
        elif len(args) == 2 and ntype == "hpeak":
            hpeak_havecontrol(args[0], args[1], npvalue, nname)
        elif len(args) == 1 and ntype == "hpeak":
            hpeak_nocontrol(args[0], npvalue, nname)
        endtime=time.time()
        print "it takes %d seconds or %d minutes or %d hours to run this program!" % (endtime-starttime,
(endtime-starttime)/60, (endtime-starttime)/3600)
```

附录 3 ChIP-seq 信号提取代码

```
#author: zjuwhw
#date: 2014-12-25
USAGE='"extract_signal_from_bigwig.py --- extract signals from bigwig file accoarding to the bed file
USAGE:
    python %s [--type=#] [--strand=INT] [--nbins=#] [--range=#] [--output=<file>] bigwigfile bedfile(or stdin/-)
#there are 4 types for this program
#type 1: the bed file is a region, output a average value appending echo line
#type 2: the bed file is a region and the nbins parameter is needed, output an aggregation plot which divide the region into nbins
#type 3: the bed file is a point and the range parameter is needed, output an aggregation plot which aggregate the signal in each position around the point
#type 4: the bed file is a point and the needed parameters are both nbins and range, output a heatmap plot
#--strand is just for type 2, 3, 4; and the INT is NUMBER of the strand column.
#input a bed formate file(the first three colnums are chr, str and end,respectively)
NOTE:
#the output value = real value/ row number of bedfile/ average tag density of bigWig file
#this program is based on the bigWigSummary tool, so the bigWigSummary must in the PATH
#if bedfile is stdin or -, the input is the stdin
#the bigwig file can be the /c/wanghw/annotation/conservation_score/genome.phastCons46way.bw when calculate conversion score.
#it is better to set the output parameter, because the bigWigSummary also output some contents into stdout'"
import os,sys,getopt,time
def ossystemresult(command):
    fp=os.popen(command,"r")
    return fp.read()
def type1(bedfp,outfp,bigwigfile):
    for line in bedfp:
        linelist=line.rstrip().split("\t")
cmd1="bigWigSummary %s %s %s %s %d" %(bigwigfile,linelist[0],linelist[1],linelist[2],1)
        score1=ossystemresult(cmd1)
        cmd2="bigWigSummary
-type=coverage %s %s %s %s %d" %(bigwigfile,linelist[0],linelist[1],linelist[2],1)
        cover=ossystemresult(cmd2)
        if score1=="n/a" or score1=="" :
            score=0
        else:
```

```
                score=float(score1)*float(cover)
            print >> outfp, "%s\t%f" % (line.rstrip(),score)
            #print cmd1,cmd2,score1,cover,score
        bedfp.close()
        outfp.close()
def type2(bedfp,outfp,bigwigfile,nbins,nstrand):
        summary=[0]*nbins
        avr_tag_density=ossystemresult("'bigWigInfo %s |awk -F " " '$1=="mean:"{print $2}' '" %
bigwigfile)
        nlines=0
        for line in bedfp:
            nlines+=1
            linelist=line.rstrip().split("\t")
cmd1="bigWigSummary %s %s %s %s %d" %(bigwigfile,linelist[0],linelist[1],linelist[2],nbins)
            score1=ossystemresult(cmd1)
            score1_list=score1.rstrip().split("\t")
            cmd2="bigWigSummary
-type=coverage %s %s %s %s %d" %(bigwigfile,linelist[0],linelist[1],linelist[2],nbins)
            cover=ossystemresult(cmd2)
            cover_list=cover.rstrip().split("\t")
            for i1 in range(nbins):
                if not nstrand:
                    i2=i1
                else:
                    if linelist[nstrand-1]=="+":
                        i2=i1
                    elif linelist[nstrand-1]=="-":
                        i2=-(i1+1)
                try:
                    score1_i=score1_list[i1]
                    cover_i=cover_list[i1]
                    summary[i2]+=float(score1_i)*float(cover_i)
                except:
                    summary[i2]+=0
        for j in range(nbins):
            print >> outfp, "%d\t%.8f" % (j+1,float(summary[j])/float(nlines)/float(avr_tag_density))
        bedfp.close()
        outfp.close()
def type3(bedfp,outfp,bigwigfile,rangelength,nstrand):
        summary=[0]*(2*rangelength+1)
```

```
    avr_tag_density=ossystemresult('"bigWigInfo %s |awk -F " " '$1=="mean:"{print $2}' '" %
bigwigfile)
    nlines=0
    for line in bedfp:
        nlines+=1
        linelist=line.rstrip().split("\t")
        if int(linelist[2])-int(linelist[1]) !=1 :
            print "type 4 need input a point bed file!!"
            sys.exit(1)
        start=int(linelist[1])-rangelength
        end=int(linelist[2])+rangelength
cmd1="bigWigSummary %s %s %d %d %d" %(bigwigfile,linelist[0],start,end,2*rangelength+1)
        score1=ossystemresult(cmd1)
        score1_list=score1.rstrip().split("\t")
        for i1 in range(2*rangelength+1):
            if not nstrand:
                i2=i1
            else:
                if linelist[nstrand-1]=="+":
                    i2=i1
                elif linelist[nstrand-1]=="-":
                    i2=-(i1+1)
            try:
                score1_i=score1_list[i1]
                summary[i2]+=float(score1_i)
            except:
                summary[i2]+=0
    #print
nlines,summary,avr_tag_density,type(nlines),type(summary),len(summary),type(avr_tag_density),float(
avr_tag_density)
    for j in range(2*rangelength+1):
        #print j-rangelength,float(summary[j])/float(nlines)/float(avr_tag_density)
        print >> outfp, "%d\t%.8f" %
(j-rangelength,float(summary[j])/float(nlines)/float(avr_tag_density))
    bedfp.close()
    outfp.close()
def type4(bedfp,outfp,bigwigfile,nbins,rangelength,nstrand):
    avr_tag_density=ossystemresult('"bigWigInfo %s |awk -F " " '$1=="mean:"{print $2}' '" %
bigwigfile)
    for line in bedfp:
```

```
            linelist=line.rstrip().split("\t")
            if int(linelist[2])-int(linelist[1]) !=1 :
                print "type 4 need input a point bed file!!"
                sys.exit(1)
            start=int(linelist[1])-rangelength
            end=int(linelist[2])+rangelength
            cmd1="bigWigSummary %s %s %d %d %d" %(bigwigfile,linelist[0],start,end,nbins)
            score1=ossystemresult(cmd1)
            score1_list=score1.rstrip().split("\t")
            cmd2="bigWigSummary
-type=coverage %s %s %s %s %d" %(bigwigfile,linelist[0],start,end,nbins)
            cover=ossystemresult(cmd2)
            cover_list=cover.rstrip().split("\t")
            summary_line=[0]*nbins
            for i1 in range(nbins):
                if not nstrand:
                    i2=i1
                else:
                    if linelist[nstrand-1]=="+":
                        i2=i1
                    elif linelist[nstrand-1]=="-":
                        i2=-(i1+1)
                try:
                    score1_i=score1_list[i1]
                    cover_i=cover_list[i1]
                    summary_line[i2]+=float(score1_i)*float(cover_i)/float(avr_tag_density)
                except:
                    summary_line[i2]+=0
                    #print type(score1_i) , type(cover_i), score1_i, cover_i
            #print summary_line,avr_tag_density
            #print "\t".join(map(str,summary_line))
            print >> outfp,"\t".join(map(str,summary_line))
        bedfp.close()
        outfp.close()
def aggragation_plot(output):
        rfilename=output+".r"
        rfilefp = open(rfilename,"w")
        print >> rfilefp,'''
        file="%s"
        data=read.table(file)
```

```
        png(paste(file,".png",sep=""))
        plot(data[,1],data[,2],type="l",main=file)
        dev.off()'" % output
        rfilefp.close()
if __name__ == '__main__':
        starttime=time.time()
        if len(sys.argv) < 2:
                print USAGE % sys.argv[0]
                sys.exit(1)
        opts, args = getopt.getopt(sys.argv[1:], "", ["type=","strand=","nbins=","range=","output="])
        if len(args)!=2:
                print "both the bigwigfile and the bedfile are needed\n"
                sys.exit(1)
        # defaults
        ntype=1
        nbins=1
        nstrand=False
        rangelength=2000
        outfp=sys.stdout
        output="name"
        for o,a in opts:
                if o == '--type':
                        ntype = int(a)
                elif o == '--strand':
                        nstrand = int(a)
                elif o == '--nbins':
                        nbins = int(a)
                elif o== '--range':
                        rangelength = int(a)
                elif o == '--output':
                        outfp = open(a, "w")
                        output = a
        bigwigfile=args[0]
        if args[1]=="-" or args[1]=="stdin":
                bedfp=sys.stdin
        else:
                bedfp=open(args[1])
        if ntype==1:
                type1(bedfp,outfp,bigwigfile)
        elif ntype==2:
```

```
        type2(bedfp,outfp,bigwigfile,nbins,nstrand)
        #aggragation_plot(output)
elif ntype==3:
        type3(bedfp,outfp,bigwigfile,rangelength,nstrand)
        #aggragation_plot(output)
elif ntype==4:
        type4(bedfp,outfp,bigwigfile,nbins,rangelength,nstrand)
else:
        print "the type parameter is needed anytime"
endtime=time.time()
print "it takes %d seconds to run this program!" % (endtime-starttime)
```

附录 4 ChIP-seq 的 motif 分析

```
#Author: zjuwhw
#date: 2014-12-25
USAGE='"motif_discovery.py---used for motif discovery of ChIP-seq datas, +/-100bp region around summit using meme-chip, homer or amd software.
USAGE:
    python %s [--tools="meme-chip","homer","amd"] [--sequence=#] [--motif_db=#] [--output=#] [--d=#] _p_s.bed optional(specific for the tools)
#input is the _p_s.bed
#defaults:
  #--motif_db /c/wanghw/motif_database/Transfect_9.2.meme
  #--output the real name of _p_s.bed file name
  #--sequence /d/database/hg19/hg19.fa.masked
  #--d 200bp, when d is "fulllength" means using amd and peak full length
  #optional:
    For meme-chip: -meme-nmotifs 5 -meme-minw 6 -meme-maxw 20
    For homer: -mask -S 20 -len 8,10,12,14 -p 4 -size 200
    For amd: -T 50 -CO 0.6 -FC 1.2
Note:
The tools, meme-chip, amd, homer and bedtools, are needed in $PATH"'
import os,sys,getopt,time
def ossystemresult(command):
    fp=os.popen(command,"r")
    return fp.read()
def getbed_centor(psfp, d):
    psfp.seek(0)
    tmpbedfp=open("tmp.bed","w")
    for line in psfp:
        linelist=line.rstrip().split("\t")
        print >> tmpbedfp, "%s\t%s\t%s\t%s" %(linelist[0],str(int(linelist[4])-1),linelist[4],linelist[3])
    tmpbedfp.close()
    cmd = '" bedtools slop -i tmp.bed -g /c/wanghw/annotation/hg19.genome -b %d | awk 'BEGIN{OFS="\\t"}{print $0,".","+"; print $0, ".", "-"}' > tmp_homer.bed'" %(int(d/2))
    os.system(cmd)
    os.remove("tmp.bed")
def getfasta_centor(psfp, d, sequence):
    psfp.seek(0)
    tmpbedfp=open("tmp.bed","w")
    for line in psfp:
```

```
            linelist=line.rstrip().split("\t")
            print >>tmpbedfp,"%s\t%s\t%s\t%s" % (linelist[0],str(int(linelist[4])-1),linelist[4],linelist[3])
    tmpbedfp.close()
    cmd="bedtools slop -i tmp.bed -g /c/wanghw/annotation/hg19.genome -b %d|bedtools getfasta -fi %s -bed - -fo tmp.fa" %(int(d/2), sequence)
    os.system(cmd)
    os.remove("tmp.bed")
def getfasta_fulllength(psfp, sequence):
    psfp.seek(0)
    tmpbedfp=open("tmp.bed","w")
    for line in psfp:
            linelist=line.rstrip().split("\t")
            print >>tmpbedfp,"%s\t%s\t%s\t%s" % (linelist[0],linelist[1],linelist[2],linelist[3])
    tmpbedfp.close()
    cmd="bedtools getfasta -fi %s -bed %s -fo tmp.fa" %(sequence, "tmp.bed")
    os.system(cmd)
    os.remove("tmp.bed")
def getfasta_region(psfp):
    psfp.seek(0)
    tmpbedfp=open("tmp.bed","w")
    for line in psfp:
            linelist=line.rstrip().split("\t")
            print >> tmpbedfp,"%s\t%s\t%s\t%s" % (linelist[0],linelist[1],linelist[2],linelist[3])
    cmd="bedtools getfasta -fi %s -bed tmp.bed -fo tmp.fa" % (sequence)
    os.system(cmd)
    os.remove("tmp.bed")
def meme_chip(name, motif_db, optional):
    cmd = "meme-chip %s -o %s -db %s tmp.fa" % (optional, name, motif_db)
    os.system(cmd)
    os.remove("tmp.fa")
def homer_findMotifsGenome(name, optional):
    cmd = "findMotifsGenome.pl tmp_homer.bed hg19 %s %s" % (name, optional)
    os.system(cmd)
    os.remove("tmp_homer.bed")
def amd(name, optional, motif_db):
    cmd = "AMD.bin %s -F tmp.fa -B /c/wanghw/software/AMD-motifjournal.pone.0024576.s004/Bgresult1000.txt " % optional
    print cmd
    os.system(cmd)
    nline = ossystemresult("cat tmp.fa|wc -l")
```

```
    nline = int(nline)
    os.rename("tmp.fa", "%s.fa" % name)
    os.rename("tmp.fa.Matrix", "%s.Matrix" % name)
    os.rename("tmp.fa.Details", "%s.Details" % name)
    matrixfp = open("%s.Matrix" % name)
    motifmemefp= open("%s.meme" % name, "w")
    motifmatrix={}
    nlinematrix={}
    switch=False
    for line in matrixfp:
        if line.rstrip().endswith(":"):
            a=line.rstrip().rstrip(":")
            nlinematrix[a]=0
            motifmatrix[a]=""
            switch=True
        elif switch == True:
            motifmatrix[a]+="\t".join(line.split("\t")[1:])
            nlinematrix[a]+=1
    print motifmatrix
    print nlinematrix
    modelhead="'MEME version 4.4
ALPHABET= ACGT
strands: + -
Background letter frequencies (from uniform background):
A 0.25000 C 0.25000 G 0.25000 T 0.25000
'"
    modelmotif="'MOTIF %s %s
letter-probability matrix: alength= 4 w= %d nsites= %d E= 0
%s
'"
    print >> motifmemefp, modelhead
    for motifname in motifmatrix.keys():
        print >> motifmemefp, modelmotif % (motifname, motifname, nlinematrix[motifname], nline,
motifmatrix[motifname])
    motifmemefp.close()
    matrixfp.close()
    cmd_tomtom = "tomtom -o %s %s %s" % (("%s_tomtom_out" % name),("%s.meme" %
name),motif_db)
    os.system(cmd_tomtom)
if __name__ == '__main__':
```

```
starttime=time.time()
if len(sys.argv) < 2:
    print USAGE % sys.argv[0]
    sys.exit(1)

# defaults
motif_db = "/c/wanghw/motif_database/Transfect_9.2.meme "
sequence = "/d/database/hg19/hg19.fa.masked"
optional_database={"homer": " -mask -S 20 -len 8,10,12,14 -p 4 -size 200", "meme-chip": "
-meme-nmotifs 5 -meme-minw 6 -meme-maxw 20", "amd": "-T 50 -CO 0.6 -FC 1.2"}
tools = "meme-chip"
d=200
opts,args=getopt.getopt(sys.argv[1:],"",["tools=","motif_db=","sequence=","output=","d="])
for o,a in opts:
    if o == '--tools':
        tools = a
        optional = optional_database[tools]
        name = os.path.basename(args[0]).replace("_p_s.bed","")+"_"+tools+"_output"
    elif o=='--motif_db':
        motif_db=a
    elif o == '--sequence' :
        sequence = a
    elif o=='--output':
        name    =a +"_"+tools+"_output"
    elif o=='--d' :
        d = a
psfp = open(args[0])
if len(args) !=1 :
    optional = " ".join(args[1:])
if tools == "meme-chip":
    d = int(d)
    getfasta_centor(psfp, d, sequence)
    meme_chip(name, motif_db, optional)
elif tools == "homer":
    d = int(d)
    getbed_centor(psfp, d)
    homer_findMotifsGenome(name, optional)
elif tools == "amd":
    if d == "fulllength":
        getfasta_fulllength(psfp, sequence)
```

```
        else:
            d = int(d)
            getfasta_centor(psfp, d, sequence)
        amd(name, optional, motif_db)
endtime=time.time()
print "it takes %d seconds to run this program!" % (endtime-starttime)
```

附录 5 microarray 流程化分析代码

```
#author: zjuwhw
#date: 2014-12-26
USAGE='"affy_array_pipeline.py --- affy microarray pipeline using R to do rma, mas5.0 and/or not customCDF normalization
USAGE:

    python %s [--affyname=#] [--output=#] [--ann_affy_path=#] [--ann_refseq_bed12=#] cel_RAW_dictionary

Default:
--output is the basename of cel_RAW_dictionary
These affy microarrays are available:
"hgu133a","hgu133a2","hgu133b","hgu133plus2","hgu219","hgu95a","hgu95av2","hgu95b","hgu95c","hgu95d","hgu95e","u133aaofav2"
--ann_affy_path is the path of affy annotation. The affy annoation can be built using another python program "affy_build_annotation.py". The default ann_affy_path is "/c/wanghw/annotation/affy/" and ends with "_ann.txt"
--ann_refseq_bed12 is the path of refseq bed12 annotation files, which could be downloaded from ucsc website. The default path is "/c/wanghw/annotation/refseq_hg19_07292013.bed""
import os,sys,getopt,time
affys = ["hgu133a","hgu133a2","hgu133b","hgu133plus2","hgu219","hgu95a","hgu95av2","hgu95b","hgu95c","hgu95d","hgu95e","u133aaofav2"]
customcdfname_refseq = {"hgu133a":"HGU133A_Hs_REFSEQ",
                        "hgu133a2":"HGU133A2_Hs_REFSEQ",
                        "hgu133b":"HGU133B_Hs_REFSEQ",
                        "hgu133plus2":"HGU133Plus2_Hs_REFSEQ",
                        "hgu219":"HGU219_Hs_REFSEQ",
                        "hgu95a":"HGU95A_Hs_REFSEQ",
                        "hgu95av2":"HGU95Av2_Hs_REFSEQ",
                        "hgu95b":"HGU95B_Hs_REFSEQ",
                        "hgu95c":"HGU95C_Hs_REFSEQ",
                        "hgu95d":"HGU95D_Hs_REFSEQ",
                        "hgu95e":"HGU95E_Hs_REFSEQ",
                        "u133aaofav2":"U133AAofAv2_Hs_REFSEQ"}
customcdfname_ensg = {"hgu133a":"HGU133A_Hs_ENSG",
                      "hgu133a2":"HGU133A2_Hs_ENSG",
                      "hgu133b":"HGU133B_Hs_ENSG",
                      "hgu133plus2":"HGU133Plus2_Hs_ENSG",
```

```
                    "hgu219":"HGU219_Hs_ENSG",
                    "hgu95a":"HGU95A_Hs_ENSG",
                    "hgu95av2":"HGU95Av2_Hs_ENSG",
                    "hgu95b":"HGU95B_Hs_ENSG",
                    "hgu95c":"HGU95C_Hs_ENSG",
                    "hgu95d":"HGU95D_Hs_ENSG",
                    "hgu95e":"HGU95E_Hs_ENSG",
                    "u133aaofav2":"U133AAofAv2_Hs_ENSG"}
def cel2txt(cel_raw_dir, affy_type, name):
    print "Step1: it's begin to do normalization ..."
    Rcode = '''
dir="%s"
library(affy)
Data=ReadAffy(celfile.path=dir)
eRMA=rma(Data)
write.exprs(eRMA,file="%s_rma.txt")
eMAS = mas5(Data)
write.exprs(eMAS, file="%s_mas.txt")
#customcdf_refseq
Data=ReadAffy(celfile.path=dir, cdfname = "%s")
eRMA=rma(Data)
write.exprs(eRMA,file="%s_rma_customCDF_refseq.txt")
eMAS = mas5(Data)
write.exprs(eMAS, file="%s_mas_customCDF_refseq.txt")
#customcdf_ensg
#Data=ReadAffy(celfile.path=dir, cdfname = "%s")
#eRMA=rma(Data)
#write.exprs(eRMA,file="%s_rma_customCDF_ensg.txt")
#eMAS = mas5(Data)
#write.exprs(eMAS, file="%s_mas_customCDF_ensg.txt")
'''
    f = open("cel2txt_%s.r" % name, "w")
    print >> f, Rcode % (cel_raw_dir, name, name, customcdfname_refseq[affy_type], name, name,
customcdfname_ensg[affy_type], name, name)
    f.close()
    cmd = "Rscript cel2txt_%s.r" % name
    os.system(cmd)
    os.remove("cel2txt_%s.r" % name)
def txt2anntxt(name, ann_refseq_bed12, ann_affy_path):
    print "Step2: it's begin to annotation the probe or refseq id using the gene symbol ..."
```

```
    cmd1 = r"' awk -F '\t' -v OFS='\t'
'BEGIN{while((getline<"%s")>0)l[$1]=$3}NR>1{$1=$1"|"l[$1]}{print $0}'    %s > %s "' %
(ann_affy_path, name + "_rma.txt", name + "_rma.ann.txt")
    cmd2 = r"' awk -F '\t' -v OFS='\t'
'BEGIN{while((getline<"%s")>0)l[$1]=$3}NR>1{$1=$1"|"l[$1]}{print $0}'    %s > %s "' %
(ann_affy_path, name + "_mas.txt", name + "_mas.ann.txt")
    cmd3 = r"' awk -F '\t' -v OFS='\t'
'BEGIN{while((getline<"%s")>0)l[$5]=$4}NR>1{split($1,x,".");if(x[1] in
l){$1=$1"|"l[x[1]]}else{$1=$1"|NA"}}{print $0}'    %s > %s "' % (ann_refseq_bed12, name +
"_rma_customCDF_refseq.txt", name + "_rma_customCDF_refseq.ann.txt" )
    cmd4 = r"' awk -F '\t' -v OFS='\t'
'BEGIN{while((getline<"%s")>0)l[$5]=$4}NR>1{split($1,x,".");if(x[1] in
l){$1=$1"|"l[x[1]]}else{$1=$1"|NA"}}{print $0}'    %s > %s "' % (ann_refseq_bed12, name +
"_mas_customCDF_refseq.txt", name + "_mas_customCDF_refseq.ann.txt" )
    #print cmd1
    #print cmd2
    #print cmd3
    #print cmd4
    os.system(cmd1)
    os.system(cmd2)
    os.system(cmd3)
    os.system(cmd4)
def rowcollapse(ann_file):
    print "Step3: it's begin to collapse the row with same gene symbol using the highest expression
probe or refseq id ..."
    cmd1 = r"'head -1 %s > %s "' % (ann_file, ann_file.replace("ann.txt","HighestUniqSymbols.txt"))
    cmd2 = r"'awk
'NR!=1{split($1,x,"|");$1=x[2];if($1!="NA"){n=0;for(i=2;i<=NF;i++){n=n+$i};print n,$0}}' %s | sort
-k2,2 -k1,1nr|awk 'BEGIN{a="A"}$2!=a{printf $2;for(i=3;i<=NF;i++){printf "\t"$i};print
"";a=$2}' >> %s "'% (ann_file, ann_file.replace("ann.txt","HighestUniqSymbols.txt"))
    #print cmd1
    #print cmd2
    os.system(cmd1)
    os.system(cmd2)
if __name__ == '__main__':
    starttime=time.time()
    if len(sys.argv) < 2:
        print USAGE % sys.argv[0]
        sys.exit(1)
```

```
    opts, args = getopt.getopt(sys.argv[1:], "",
["affyname=","output=","ann_affy_path=","ann_refseq_bed12="])
    # defaults
    ann_refseq_bed12 = "/c/wanghw/annotation/refseq_hg19_07292013.bed"
    cel_raw_dir = args[0].rstrip("/")
    nname = os.path.basename(cel_raw_dir)
    for o,a in opts:
        if o == '--affyname':
            naffyname = a
            ann_affy_path = "/c/wanghw/annotation/affy/%s_ann.txt" % naffyname
        elif o == '--output':
            nname = a
        elif o == '--ann_affy_path':
            ann_affy_path = a
        elif o == '--ann_refseq_bed12':
            ann_refseq_bed12 = a
    if naffyname not in affys:
        print "The %s is not in the list as below:" % naffyname
        print ", ".join(affys)
        sys.exit(1)
    cel2txt(cel_raw_dir, naffyname, nname)
    txt2anntxt(nname, ann_refseq_bed12, ann_affy_path)
    rowcollapse(nname + "_mas.ann.txt" )
    rowcollapse(nname + "_rma.ann.txt" )
    rowcollapse(nname + "_mas_customCDF_refseq.ann.txt" )
    rowcollapse(nname + "_rma_customCDF_refseq.ann.txt" )
    os.remove("%s_rma.txt" % nname)
    os.remove("%s_mas.txt" % nname)
    os.remove("%s_rma_customCDF_refseq.txt" % nname)
    os.remove("%s_mas_customCDF_refseq.txt" % nname)
    #os.remove("%s_rma_customCDF_ensg.txt" % nname)
    #os.remove("%s_mas_customCDF_ensg.txt" % nname)
    endtime=time.time()
    print "it takes %d seconds or %d minutes or %d hours to run this program!" % (endtime-starttime,
(endtime-starttime)/60, (endtime-starttime)/3600)
```

附录 6 Kasumi-1 细胞中 AML1/ETO 抑制的基因

Genename	GFOLD	Log2 of fold change	con RPKM	shAE RPKM
C11orf21	6.46	7.10	0.02	2.16
BPI	6.20	6.28	1.55	116.81
PRG2	6.18	6.22	14.43	1049.83
SELPLG	5.46	5.67	0.15	7.54
CST7	5.52	5.65	1.26	61.60
PRG3	5.16	5.40	0.33	13.48
CEACAM6	5.07	5.26	0.20	7.52
LAPTM5	4.90	4.95	2.90	87.20
PRAM1	4.24	4.51	0.11	2.46
RNASE3	4.24	4.48	0.41	9.07
FBN1	4.15	4.27	0.12	2.20
S100A8	3.83	4.24	0.19	3.48
NKG7	3.83	4.13	0.22	3.83
NPAS1	3.84	4.09	0.14	2.26
FBP1	3.70	4.03	0.09	1.52
CLEC12A	3.76	3.99	0.23	3.57
ANXA1	3.84	3.96	0.81	12.33
CEACAM4	3.54	3.95	0.08	1.24
PLB1	3.64	3.84	0.09	1.27
EPX	3.68	3.75	1.46	19.10
TARP	3.46	3.66	0.45	5.57
RASGRP2	3.56	3.61	3.01	35.83
GAPDHS	3.44	3.55	1.05	11.98
MZB1	3.28	3.54	0.31	3.51
ECRP	3.35	3.51	1.18	13.17
CTSG	3.45	3.48	20.38	222.00
PLCB2	3.28	3.48	0.10	1.09
AP5B1	3.42	3.46	2.48	26.53
SYTL1	3.27	3.34	2.10	20.68
C1orf162	3.09	3.28	0.57	5.43
SPNS3	3.13	3.27	0.48	4.55
RMND5B	3.04	3.25	0.23	2.17
MGLL	2.94	3.12	0.15	1.25
HCST	2.95	3.11	1.54	12.98
DYNAP	2.81	3.09	0.12	0.99
SLA	3.01	3.08	1.25	10.31
RWDD2A	2.73	3.03	0.15	1.20

CD24	3.00	3.02	22.15	175.07
MFAP4	2.89	2.96	2.08	15.73
BANK1	2.74	2.94	0.14	1.07
PIWIL4	2.84	2.94	0.76	5.63
IGLL5	2.81	2.91	2.07	15.13
RASSF5	2.73	2.89	0.22	1.56
ANXA2R	2.70	2.86	0.61	4.30
ANXA3	2.56	2.84	0.15	1.07
CDH3	2.77	2.83	1.21	8.34
OGG1	2.47	2.76	0.14	0.93
FBLN5	2.56	2.74	0.23	1.49
LGALS1	2.53	2.74	0.83	5.40
VAV1	2.66	2.73	1.74	11.24
CCL4L1	2.44	2.70	0.21	1.36
CCL4L2	2.44	2.70	0.21	1.36
RAB37	2.60	2.64	3.52	21.44
ALDH3B1	2.52	2.59	1.81	10.55
AVPR2	2.29	2.54	0.20	1.15
UGT3A2	2.44	2.51	1.60	8.91
RNASE2	2.47	2.51	19.98	110.94
NFE2	2.39	2.51	0.84	4.65
ARHGAP4	2.44	2.47	5.66	30.59
C19orf77	2.28	2.44	0.99	5.23
GSE1	2.38	2.43	1.16	6.05
PSD4	2.36	2.42	1.17	6.08
CRIP1	1.92	2.40	0.17	0.91
SEL1L3	2.25	2.40	0.20	1.04
P2RY2	2.33	2.36	3.37	16.85
CEBPE	2.33	2.36	22.72	113.24
ENTPD7	2.31	2.35	1.84	9.10
PPP1R27	2.22	2.34	2.05	10.09
LOC255130	2.18	2.34	0.67	3.29
TRIB1	2.24	2.31	1.31	6.32
SPI1	2.24	2.27	14.47	68.18
GCA	2.20	2.25	2.87	13.32
PSTPIP1	2.07	2.25	0.36	1.68
RASAL3	2.16	2.25	0.89	4.11
IL6R	2.10	2.24	0.19	0.88
RAC2	2.16	2.19	16.82	74.79

CCDC88B	2.12	2.19	0.99	4.36
TP53TG1	2.02	2.17	1.33	5.84
DHRS9	2.02	2.16	0.82	3.56
RASSF2	2.09	2.14	1.76	7.58
ZCCHC24	2.02	2.14	0.34	1.46
PTPN6	2.10	2.13	6.59	28.15
SLC16A3	1.91	2.12	0.22	0.95
APOBR	1.97	2.12	0.25	1.06
TBC1D10C	2.04	2.11	2.83	11.91
IGFBP7	2.06	2.10	11.26	47.10
MIR223	1.98	2.10	14.20	59.26
MCU	2.06	2.09	7.66	31.80
RHOXF2	2.05	2.07	53.68	219.47
RHOXF2B	2.05	2.07	53.36	217.98
KCNAB2	2.03	2.06	7.78	31.53
CXCR4	1.98	2.05	2.34	9.48
GBGT1	1.84	2.05	0.28	1.12
CEBPD	1.96	2.04	3.23	12.91
PPM1M	1.89	2.03	0.56	2.24
NINJ2	1.89	2.00	1.55	6.05
P4HB	1.99	2.00	531.81	2068.53
SLC43A3	1.96	1.99	10.12	39.26
LGALS12	1.85	1.98	0.83	3.17
HIST1H2AC	1.66	1.96	0.42	1.62
SIPA1	1.90	1.96	1.94	7.33
LINC00884	1.73	1.95	0.39	1.46
NINJ1	1.78	1.94	0.66	2.48
SLC51A	1.85	1.94	1.92	7.16
MYO1F	1.89	1.93	3.41	12.66
MBP	1.85	1.92	0.92	3.41
CHST12	1.82	1.92	1.13	4.15
UNC13D	1.88	1.91	5.42	19.82
KLF7	1.79	1.90	0.24	0.86
TTN	1.87	1.90	0.33	1.19
TMEM187	1.70	1.89	0.37	1.35
PTK2B	1.84	1.89	2.58	9.33
SIRPB1	1.77	1.88	0.80	2.87
CYFIP2	1.77	1.87	0.32	1.15
GPR97	1.70	1.87	0.30	1.09

S100P	1.77	1.87	4.22	15.03
MIR612	1.26	1.87	0.56	2.03
PRRT4	1.81	1.87	3.76	13.36
NIPAL2	1.69	1.86	0.32	1.13
SLC24A6	1.82	1.86	4.39	15.53
NCF2	1.77	1.86	1.40	4.94
CROT	1.75	1.85	0.73	2.57
ATG16L2	1.79	1.84	4.20	14.62
PLAC8	1.79	1.83	7.27	25.23
COL23A1	1.78	1.83	3.69	12.79
FES	1.77	1.81	5.71	19.56
LPCAT2	1.78	1.81	7.16	24.39
ME3	1.66	1.80	0.53	1.82
LOC100129858	1.62	1.80	0.36	1.23
S100A6	1.59	1.80	0.79	2.68
ERN1	1.67	1.80	0.36	1.22
C3AR1	1.74	1.79	3.44	11.60
ARHGEF10	1.70	1.79	0.49	1.65
LOC153684	1.61	1.78	0.39	1.29
CD82	1.73	1.77	7.71	25.66
SCARNA2	1.52	1.77	0.85	2.84
LDHD	1.62	1.77	0.54	1.79
C5AR2	1.59	1.77	0.58	1.92
MIR663B	1.06	1.76	0.36	1.22
HSD17B10	1.71	1.76	9.95	32.77
ABHD8	1.67	1.75	1.52	5.00
EAF2	1.69	1.75	6.23	20.43
SRGN	1.72	1.74	35.03	114.09
SH3BGRL3	1.70	1.74	17.33	56.35
HIST2H2AA4	1.58	1.73	1.72	5.58
HIST2H2AA3	1.58	1.73	1.72	5.58
TTN-AS1	1.64	1.72	1.71	5.49
ID2	1.56	1.71	0.73	2.34
FUT4	1.67	1.69	8.21	25.87
SNORD4B	1.46	1.69	5.72	18.01
HSPA6	1.59	1.69	1.03	3.22
ANGPTL6	1.59	1.68	1.42	4.45
GLRB	1.57	1.68	0.66	2.06
S100A4	1.63	1.68	18.82	58.69

SLC48A1	1.59	1.67	1.17	3.63
ADRBK1	1.65	1.66	35.83	110.25
SLC31A2	1.56	1.65	1.51	4.63
GNA15	1.60	1.65	4.36	13.33
FLOT2	1.60	1.65	4.37	13.33
C9orf9	1.44	1.65	0.80	2.44
PTPN22	1.57	1.64	1.18	3.58
RAB24	1.55	1.64	1.79	5.42
HSPA7	1.56	1.62	2.77	8.30
PYCARD	1.54	1.62	4.38	13.13
PLEC	1.58	1.62	1.02	3.07
CERCAM	1.58	1.62	6.82	20.38
PLP2	1.52	1.61	2.56	7.61
ITM2C	1.57	1.61	9.76	29.01
SBF2-AS1	1.50	1.61	0.75	2.22
MARCKS	1.48	1.60	0.39	1.16
TRPC2	1.38	1.58	0.50	1.47
BZRAP1-AS1	1.52	1.58	3.12	9.07
C16orf98	1.44	1.58	1.20	3.50
DYNLRB1	1.52	1.57	15.89	45.80
NUDT18	1.41	1.56	0.70	2.03
SLC15A2	1.50	1.56	1.39	3.98
TMEM53	1.41	1.55	0.71	2.03
RIPK3	1.35	1.55	0.32	0.93
LOC541473	1.46	1.55	3.19	9.09
ABHD16B	1.39	1.55	0.52	1.49
DENND1A	1.50	1.55	1.99	5.67
ANXA5	1.50	1.54	9.76	27.62
UNC93B1	1.47	1.53	2.66	7.48
PCSK4	1.37	1.53	0.35	0.99
FKBP6	1.40	1.53	1.03	2.89
LRP5L	1.40	1.53	0.90	2.53
PNP	1.50	1.52	21.53	60.24
RASSF7	1.38	1.52	0.78	2.16
AMPD3	1.48	1.52	3.38	9.43
CITED4	1.41	1.52	1.72	4.79
AMT	1.42	1.51	1.31	3.64
MROH6	1.45	1.51	2.14	5.94
ALAS1	1.48	1.51	20.66	57.17

PARVG	1.46	1.50	4.15	11.46
GLRX	1.44	1.50	6.90	19.03
LACTB	1.41	1.49	1.74	4.77
TNFAIP8L2	1.38	1.49	1.68	4.59
ARHGEF17	1.41	1.49	0.47	1.29
GLA	1.45	1.48	16.76	45.52
ANKRD13D	1.43	1.48	5.84	15.84
AGTRAP	1.44	1.48	17.49	47.39
SEC24D	1.44	1.47	5.66	15.29
CXorf21	1.37	1.47	1.33	3.59
ANPEP	1.45	1.47	22.29	59.96
ACTN1	1.44	1.46	17.46	46.83
HIST1H2BN	1.10	1.46	0.42	1.13
RGCC	1.39	1.45	6.45	17.16
HP07349	1.13	1.45	0.53	1.42
HAGHL	1.30	1.44	0.98	2.58
SLC28A3	1.38	1.44	1.51	3.97
FERMT3	1.39	1.43	7.36	19.31
OSBPL5	1.36	1.42	1.88	4.90
STAG3L3	1.39	1.42	28.56	74.20
ATP10D	1.32	1.41	0.48	1.25
MIR503HG	1.29	1.41	2.32	5.99
BCL6	1.24	1.40	0.29	0.76
FCER1G	1.22	1.40	1.33	3.43
GSN	1.37	1.40	12.26	31.52
PCED1B-AS1	1.22	1.40	0.30	0.77
SMPD1	1.30	1.40	1.00	2.57
TOM1	1.32	1.39	2.38	6.08
CYBA	1.35	1.38	43.98	111.42
GHRL	1.14	1.38	0.40	1.01
STAG3L1	1.36	1.38	27.14	68.69
LGALS9B	1.24	1.38	1.07	2.72
ALOX5AP	1.23	1.38	1.36	3.44
LRMP	1.33	1.37	4.12	10.42
GRK6	1.32	1.37	5.08	12.80
IGLL3P	1.31	1.37	10.88	27.42
BBS2	1.31	1.37	2.05	5.16
RNASEH2C	1.34	1.37	10.60	26.65
LGALS9C	1.26	1.37	1.28	3.21

AZU1	1.36	1.37	717.43	1802.46
SFXN5	1.29	1.37	1.02	2.55
RHPN1	1.22	1.37	0.33	0.82
PTPN12	1.25	1.37	0.59	1.47
C7orf41	1.30	1.37	1.01	2.54
TOB1-AS1	1.16	1.36	0.46	1.15
OPRL1	1.22	1.36	0.40	1.00
ADAMTS20	1.27	1.36	0.56	1.40
SLC22A5	1.21	1.35	0.36	0.91
RMRP	1.17	1.35	2.88	7.16
ARHGEF18	1.32	1.35	4.00	9.94
PVRL1	1.10	1.35	0.29	0.72
NAPRT1	1.30	1.35	6.08	15.05
C11orf71	1.17	1.34	1.20	2.95
GCHFR	1.18	1.34	1.50	3.70
TCIRG1	1.28	1.34	2.84	6.99
MIR142	1.12	1.33	6.54	16.05
VSTM1	1.24	1.33	3.62	8.85
TMEM102	1.23	1.33	1.29	3.16
TRIM74	1.19	1.32	1.17	2.84
TRIM73	1.19	1.32	1.17	2.84
BCL3	1.15	1.32	0.46	1.12
LINGO3	1.19	1.32	0.76	1.86
RPL13AP6	1.25	1.32	8.35	20.26
SYK	1.26	1.31	2.39	5.76
BLVRB	1.19	1.30	2.21	5.32
CHRFAM7A	1.21	1.30	0.99	2.37
PGM3	1.22	1.30	1.82	4.38
TRPM4	1.22	1.30	1.13	2.70
SNORD33	1.20	1.30	31.95	76.60
LOC642852	1.22	1.30	0.58	1.40
RAB5B	1.28	1.29	22.17	52.89
CXCR2	1.13	1.29	0.38	0.90
KCNH2	1.24	1.29	4.66	11.09
NDRG1	1.22	1.29	1.93	4.60
FCGR2A	1.24	1.29	4.32	10.26
PIGV	1.23	1.29	3.17	7.52
STAG3L2	1.26	1.29	29.30	69.56
FNBP1	1.25	1.29	3.87	9.17

FCGR2C	1.19	1.28	2.22	5.27
ELL2	1.20	1.28	0.64	1.50
SPN	1.26	1.28	12.92	30.55
FAM173A	1.14	1.28	1.42	3.35
EFHD2	1.23	1.27	5.57	13.09
ADAM8	1.15	1.27	0.57	1.35
AIG1	1.20	1.27	3.30	7.74
VAT1	1.25	1.27	36.96	86.81
RGS14	1.18	1.27	1.45	3.41
PRSS57	1.25	1.26	228.67	534.74
SNX20	1.17	1.26	2.06	4.80
BHLHE40	1.11	1.26	0.39	0.92
FAM214B	1.16	1.26	0.86	2.01
FLNA	1.24	1.26	8.45	19.66
ULK4P3	1.13	1.25	1.44	3.35
MFSD3	1.19	1.25	5.01	11.60
ULK4P1	1.12	1.25	1.72	3.99
ULK4P2	1.12	1.25	1.72	3.99
TFAP2A	1.11	1.25	0.41	0.95
C16orf58	1.21	1.24	6.22	14.34
SLCO3A1	1.11	1.24	0.52	1.21
ARHGAP9	1.20	1.24	8.12	18.73
CCT6B	1.08	1.24	0.58	1.33
CCNA1	1.04	1.24	0.33	0.75
SLC25A45	1.14	1.24	0.91	2.09
GMPPA	1.15	1.24	2.39	5.48
ELANE	1.23	1.23	557.50	1277.57
HMHA1	1.20	1.23	5.42	12.42
SAT1	1.20	1.23	24.45	55.97
FAM98C	1.10	1.23	1.17	2.68
MYO18B	1.19	1.23	2.31	5.25
ZDHHC24	1.16	1.23	4.74	10.80
KLF10	1.12	1.22	0.76	1.73
OS9	1.21	1.22	48.53	110.12
INO80B	1.14	1.22	3.38	7.66
PRTN3	1.21	1.22	2347.17	5315.23
GMPPB	1.15	1.21	3.52	7.96
DNAJC5	1.19	1.21	11.10	25.02
CAT	1.20	1.21	53.47	120.54

ANKRD33B	1.17	1.21	1.68	3.78
LOC100128822	1.07	1.21	0.96	2.16
TNFAIP8	1.18	1.21	18.33	41.23
SLC5A2	1.13	1.21	1.79	4.02
CPT1A	1.18	1.21	10.57	23.72
RNPEPL1	1.15	1.20	2.75	6.17
RIN3	1.16	1.20	3.92	8.79
CAPN3	1.05	1.20	0.70	1.56
AIF1	1.13	1.20	1.61	3.61
LINC00926	1.14	1.20	3.13	7.00
GSN-AS1	1.08	1.20	0.40	0.90
GABARAPL1	1.11	1.20	1.90	4.24
TLE4	1.13	1.20	1.20	2.68
GAB3	1.09	1.20	0.52	1.16
KIF21B	1.13	1.19	1.03	2.29
CARD9	1.05	1.19	0.70	1.56
IDUA	1.07	1.19	0.94	2.09
CCDC159	1.05	1.18	1.37	3.03
EXD3	1.05	1.18	0.56	1.24
CTSD	1.16	1.18	40.48	89.20
NOTCH2	1.04	1.18	0.34	0.76
MXRA7	1.07	1.18	1.18	2.60
SLC39A11	1.15	1.18	13.22	29.10
ABAT	1.05	1.18	0.32	0.71
CALR	1.17	1.18	1701.31	3741.82
HDAC4	1.13	1.18	1.33	2.93
CD244	1.11	1.17	2.78	6.11
LRFN4	1.11	1.17	2.64	5.79
LGALS9	1.14	1.17	20.71	45.47
TOR2A	1.11	1.17	5.41	11.85
ABCA3	1.07	1.17	0.46	1.01
IGLL1	1.15	1.16	188.41	411.15
EXOSC4	1.11	1.16	9.91	21.61
CHRNA7	1.07	1.16	0.93	2.03
C16orf93	1.09	1.16	3.02	6.57
TMEM205	1.08	1.16	5.39	11.70
NFAM1	1.12	1.16	3.35	7.27
MID1IP1	1.12	1.15	12.36	26.75
SPEF2	1.07	1.15	0.67	1.44

TESK2	1.07	1.15	1.49	3.21
GMIP	1.10	1.14	4.96	10.64
NLRP3	1.00	1.14	0.32	0.69
ABCA2	1.06	1.14	0.58	1.24
RNH1	1.12	1.14	29.70	63.64
FMNL1	1.11	1.14	8.32	17.83
RASSF4	1.07	1.13	2.48	5.30
MIR1184-1	1.05	1.13	19.12	40.83
MIR1184-2	1.05	1.13	19.12	40.83
MIR1184-3	1.05	1.13	19.12	40.83
PTPRC	1.08	1.13	1.85	3.95
LOC100129550	1.05	1.13	0.76	1.62
SLC2A5	1.07	1.12	3.52	7.47
THBS3	1.05	1.12	2.06	4.36
UCP2	1.09	1.12	21.44	45.38
TSKS	1.04	1.12	2.26	4.78
DERL3	1.03	1.12	3.25	6.88
ABCA7	1.06	1.12	1.28	2.70
RAB32	1.08	1.12	12.35	26.09
RABAC1	1.06	1.12	10.17	21.47
S1PR4	1.08	1.12	12.92	27.26
PLBD2	1.05	1.11	2.67	5.61
PLD4	1.03	1.11	2.42	5.08
ARMC5	1.03	1.10	1.76	3.67
ZBTB16	1.00	1.09	1.63	3.37
ARID5A	1.03	1.09	3.38	6.99
PSMB10	1.03	1.08	7.65	15.80
GPT2	1.04	1.08	3.60	7.43
G6PD	1.05	1.08	12.38	25.54
LOC100506844	1.03	1.08	16.26	33.53
MLC1	1.07	1.08	47.28	97.31
COX15	1.04	1.08	2.91	5.97
CCDC22	1.02	1.08	3.64	7.48
ARHGEF1	1.05	1.07	10.47	21.47
TAPBP	1.03	1.07	1.31	2.67
MYADM	1.04	1.07	11.62	23.75
ALKBH7	1.01	1.07	6.88	14.04
ORAI3	1.01	1.07	3.80	7.75
F8A2	1.02	1.07	5.24	10.68

F8A3	1.02	1.07	5.24	10.68
DGKG	1.02	1.06	3.24	6.57
FBRSL1	1.03	1.06	9.76	19.79
SDF2L1	1.02	1.06	27.80	56.36
EDEM1	1.04	1.06	12.95	26.26
VAMP8	1.04	1.06	101.79	206.07
CTSA	1.03	1.06	17.22	34.84
ELF4	1.00	1.05	2.78	5.63
RNF44	1.02	1.05	7.72	15.59
SUCO	1.03	1.05	10.70	21.60
MPO	1.05	1.05	4304.99	8665.40
MYO1G	1.03	1.05	23.71	47.69
MNDA	1.01	1.05	10.33	20.76
M1AP	1.00	1.04	6.24	12.53
DOK3	1.01	1.04	11.20	22.44
ARHGAP1	1.00	1.04	6.15	12.30
IMPA2	1.00	1.04	19.50	38.90
CITED2	1.02	1.03	80.13	159.27
IRF2BP2	1.00	1.02	27.27	53.74

GFOLD (generalized fold change), a reliable statistics for expresson changes based on the posterior distribution of log fold change given by the tool GFOLD.

附录 7 Kasumi-1 细胞中 AML1/ETO 激活的基因

Genename	GFOLD	Log2 of fold change	con RPKM	shAE RPKM
TM4SF1	-5.33	-5.87	2.01	0.03
FLT1	-5.17	-5.87	1.02	0.02
DUSP27	-4.75	-5.08	1.48	0.04
ADAM28	-4.67	-4.86	4.62	0.15
PADI3	-4.46	-4.66	3.76	0.14
NPTX1	-4.37	-4.52	3.78	0.16
PDZD2	-4.31	-4.47	1.42	0.06
PTRF	-4.23	-4.33	12.01	0.58
SULT1C2	-4.02	-4.20	3.89	0.21
CDKN1A	-3.63	-3.88	2.14	0.14
BMX	-3.74	-3.84	10.84	0.74
FPR2	-3.56	-3.74	4.18	0.30
ST8SIA6	-3.31	-3.71	1.06	0.08
GLYATL2	-3.53	-3.71	5.91	0.44
GAD1	-3.53	-3.67	3.49	0.27
ANXA8	-3.46	-3.64	4.17	0.32
ANXA8L2	-3.45	-3.62	4.47	0.35
ANXA8L1	-3.45	-3.62	4.32	0.34
SLC44A2	-3.30	-3.41	5.73	0.52
THSD1	-3.33	-3.41	13.17	1.21
NR5A2	-3.29	-3.39	4.81	0.45
LOC100288570	-3.11	-3.38	3.26	0.30
SLC2A3	-3.27	-3.35	8.81	0.84
MYCN	-3.18	-3.35	3.12	0.30
LOC440895	-3.06	-3.34	3.37	0.32
CD226	-3.17	-3.32	3.75	0.37
LOC100507334	-2.91	-3.22	2.14	0.22
SLC2A14	-2.92	-3.13	1.83	0.20
SHANK3	-2.99	-3.09	2.70	0.31
TSHZ3	-2.99	-3.08	5.04	0.58
ZNF521	-2.94	-3.05	3.02	0.35
CACNB4	-2.91	-2.99	3.38	0.41
ST18	-2.88	-2.97	3.24	0.40
ADRA2A	-2.77	-2.96	1.15	0.14
FSD1	-2.74	-2.93	2.45	0.31
PLXNB2	-2.77	-2.93	1.05	0.13
BAALC	-2.83	-2.90	10.93	1.42

ARID5B	-2.74	-2.84	2.07	0.28
SLC26A9	-2.70	-2.84	1.59	0.22
RCBTB2	-2.77	-2.82	17.71	2.44
C10orf114	-2.60	-2.75	3.32	0.48
RAB27B	-2.61	-2.69	2.99	0.45
MYRF	-2.53	-2.66	1.37	0.21
LINC00085	-2.41	-2.64	1.30	0.20
BAI1	-2.48	-2.62	1.30	0.21
TMEFF1	-2.48	-2.57	7.76	1.28
TSPAN18	-2.43	-2.54	2.43	0.41
CD48	-2.33	-2.53	2.87	0.48
DOK4	-2.40	-2.51	3.72	0.63
GPR124	-2.40	-2.47	3.78	0.66
SELL	-2.32	-2.46	2.68	0.47
SHANK1	-2.28	-2.41	1.20	0.22
TMEM154	-2.31	-2.40	4.38	0.81
TPSAB1	-2.29	-2.40	8.25	1.52
GNG2	-2.33	-2.40	6.43	1.19
SLC45A3	-2.32	-2.39	8.31	1.55
HPSE	-2.30	-2.38	3.85	0.72
KCTD12	-2.29	-2.38	2.40	0.45
WASF1	-2.30	-2.37	7.24	1.36
CSRP2	-2.04	-2.36	1.19	0.22
BIN1	-2.23	-2.36	3.75	0.71
CCNJL	-2.17	-2.30	1.96	0.39
SPINK4	-1.99	-2.30	3.15	0.62
GPR84	-2.10	-2.29	2.08	0.41
FAM171A1	-2.25	-2.28	29.20	5.84
MAGED4	-2.21	-2.27	9.31	1.87
MAGED4B	-2.21	-2.27	9.31	1.87
ITPKA	-2.16	-2.27	4.63	0.93
SMAGP	-2.05	-2.26	2.09	0.42
FLI1-AS1	-1.98	-2.26	1.01	0.20
PRTFDC1	-2.12	-2.25	3.62	0.74
LRRC8C	-2.20	-2.24	7.56	1.56
SNORA11D	-1.44	-2.24	1.17	0.23
SNORA11E	-1.44	-2.24	1.17	0.23
CD53	-2.18	-2.24	21.93	4.52
NOG	-2.14	-2.24	6.97	1.44

FOSL2	-2.14	-2.22	4.81	1.01
MTSS1	-2.14	-2.19	9.71	2.07
ECM1	-2.08	-2.16	9.57	2.09
TPSD1	-2.00	-2.15	4.91	1.07
TPSB2	-2.06	-2.15	10.60	2.32
DLG5	-2.06	-2.14	1.96	0.43
ARHGEF3	-2.03	-2.11	4.35	0.98
PAG1	-2.02	-2.11	1.18	0.26
C1orf186	-2.05	-2.07	111.05	25.71
ANKRD22	-1.99	-2.06	5.17	1.21
CD34	-2.02	-2.05	42.78	10.03
LRRC70	-1.87	-2.02	1.85	0.44
STYK1	-1.86	-2.01	1.41	0.34
CTTN	-1.88	-1.99	2.37	0.58
CD69	-1.94	-1.99	19.81	4.85
LRP4	-1.89	-1.99	1.25	0.31
ADAMTS3	-1.92	-1.98	4.27	1.05
GIMAP6	-1.89	-1.97	4.51	1.12
LYN	-1.92	-1.97	11.02	2.75
MMP25	-1.91	-1.96	8.72	2.18
LIMS3-LOC440895	-1.90	-1.96	10.24	2.56
MURC	-1.79	-1.93	1.38	0.35
MDFI	-1.76	-1.91	2.68	0.70
NPR3	-1.81	-1.90	1.72	0.45
CDH26	-1.73	-1.90	2.61	0.68
TNNT1	-1.69	-1.90	1.98	0.52
ESAM	-1.82	-1.89	10.20	2.69
SIGLEC12	-1.80	-1.88	7.53	2.00
PBK	-1.79	-1.87	8.93	2.38
STAB1	-1.76	-1.86	1.10	0.30
DEPTOR	-1.81	-1.85	16.17	4.36
IL17RE	-1.74	-1.85	2.55	0.69
PLCH1	-1.77	-1.85	2.26	0.61
JAM3	-1.80	-1.84	12.90	3.50
ETV5	-1.78	-1.84	6.03	1.64
STAR	-1.76	-1.84	5.51	1.50
SCUBE1	-1.70	-1.82	1.53	0.42
GYG2	-1.64	-1.82	0.94	0.26

USP44	-1.70	-1.81	1.79	0.50
ZNF385A	-1.68	-1.81	2.44	0.68
MFSD6	-1.65	-1.80	0.81	0.23
MAP3K1	-1.76	-1.80	9.26	2.59
PMP22	-1.63	-1.79	2.01	0.56
LIMS3	-1.75	-1.79	17.78	5.00
LIMS3L	-1.75	-1.79	17.78	5.00
SIGLEC5	-1.64	-1.79	1.47	0.41
CDH2	-1.74	-1.79	7.12	2.01
VAV3	-1.74	-1.79	16.39	4.63
RNF175	-1.59	-1.77	1.58	0.45
IFI16	-1.74	-1.77	42.14	12.03
ITGAV	-1.72	-1.76	6.80	1.95
RANBP2	-1.74	-1.75	68.80	19.86
PRDM8	-1.72	-1.75	33.25	9.60
TRIM49D2P	-1.59	-1.74	2.25	0.66
MSANTD3-TMEFF1	-1.68	-1.74	9.39	2.74
GPR141	-1.56	-1.73	3.12	0.91
TRIM49D1	-1.58	-1.73	2.44	0.72
EPHX4	-1.50	-1.71	1.26	0.37
SLC18A2	-1.69	-1.71	43.20	12.84
ENC1	-1.67	-1.71	8.67	2.58
SIPA1L1	-1.63	-1.71	2.28	0.68
DMWD	-1.58	-1.69	2.11	0.64
ASPH	-1.64	-1.69	7.55	2.28
CDC42EP3	-1.59	-1.68	1.88	0.57
TRIM71	-1.60	-1.68	3.57	1.08
CR1	-1.61	-1.68	2.37	0.72
CLVS1	-1.55	-1.68	1.36	0.41
LIN28B	-1.62	-1.67	6.70	2.05
NCS1	-1.52	-1.66	0.86	0.26
PRNP	-1.62	-1.66	16.39	5.05
MMP11	-1.51	-1.66	1.71	0.53
SMO	-1.59	-1.64	7.84	2.44
SV2A	-1.59	-1.64	13.49	4.21
MIR146A	-1.04	-1.64	2.03	0.62
CNST	-1.57	-1.62	5.25	1.66
FLVCR2	-1.46	-1.62	1.10	0.35

SERPINE2	-1.41	-1.61	0.83	0.26
DMPK	-1.51	-1.61	2.94	0.94
FHL1	-1.50	-1.60	2.86	0.92
ABCD3	-1.55	-1.59	10.71	3.46
PHLDB1	-1.49	-1.59	1.54	0.50
SIPA1L2	-1.52	-1.59	2.35	0.76
TMTC2	-1.52	-1.59	3.19	1.03
HIST1H2BH	-1.13	-1.59	0.83	0.27
CKB	-1.56	-1.58	97.15	31.60
PTGER4P2-CDK2AP2P2	-1.46	-1.58	1.49	0.48
VSIG10	-1.49	-1.58	1.96	0.64
ZNF792	-1.48	-1.58	1.94	0.63
DUSP6	-1.54	-1.58	23.33	7.61
NEK7	-1.53	-1.58	8.42	2.75
LMNB1	-1.55	-1.58	62.26	20.35
ATP9A	-1.48	-1.57	1.24	0.40
SAMSN1	-1.50	-1.57	8.11	2.66
CALHM2	-1.45	-1.57	2.67	0.88
GALNT1	-1.55	-1.56	83.39	27.47
PROCR	-1.40	-1.56	1.89	0.62
PAPSS2	-1.48	-1.56	3.02	1.00
AGT	-1.40	-1.56	1.11	0.37
OLFML2A	-1.44	-1.55	0.89	0.30
CD109	-1.47	-1.55	1.41	0.47
IL32	-1.42	-1.54	6.05	2.03
LIMS1	-1.51	-1.53	36.01	12.14
KIAA1551	-1.48	-1.52	7.66	2.59
SLA2	-1.47	-1.52	13.80	4.68
FAM117A	-1.48	-1.51	23.86	8.14
TNFSF13B	-1.49	-1.51	66.74	22.80
NQO1	-1.45	-1.51	10.56	3.62
STK32B	-1.40	-1.50	2.16	0.74
RGPD2	-1.47	-1.49	19.97	6.91
ANKRD65	-1.39	-1.49	5.01	1.74
TRIM47	-1.41	-1.48	5.31	1.85
ASIC1	-1.44	-1.48	9.57	3.33
SLC2A9	-1.30	-1.48	1.22	0.43
RNF157-AS1	-1.33	-1.48	1.45	0.50

PLEKHO1	-1.43	-1.48	18.89	6.59
RGPD1	-1.45	-1.47	16.64	5.83
B3GALNT1	-1.42	-1.47	8.55	3.00
DLL3	-1.41	-1.47	9.30	3.26
MEF2C	-1.41	-1.47	2.94	1.03
DCBLD2	-1.41	-1.47	3.62	1.27
FGF16	-1.23	-1.47	3.58	1.26
APOBEC3G	-1.36	-1.46	4.04	1.43
TNIK	-1.33	-1.45	0.92	0.33
GATA2	-1.40	-1.45	9.21	3.28
FAM69B	-1.39	-1.45	8.78	3.13
NLRC5	-1.39	-1.44	5.21	1.87
PHTF2	-1.39	-1.43	5.94	2.14
SGPP1	-1.31	-1.43	1.31	0.47
CD93	-1.40	-1.43	10.30	3.72
TRAF5	-1.32	-1.43	1.66	0.60
KCTD17	-1.26	-1.42	1.50	0.54
FNBP1L	-1.36	-1.42	4.87	1.77
TNFRSF10D	-1.35	-1.42	3.77	1.37
MFSD2A	-1.27	-1.42	1.55	0.56
IGFBP4	-1.40	-1.41	100.79	36.83
POMP	-1.36	-1.41	15.42	5.64
MLKL	-1.34	-1.41	5.94	2.18
IER5L	-1.26	-1.41	1.24	0.46
ABI2	-1.36	-1.40	5.93	2.18
STT3B	-1.37	-1.39	120.91	44.95
ZNF467	-1.21	-1.38	0.97	0.36
GNAI1	-1.29	-1.38	2.38	0.89
RGPD3	-1.36	-1.38	29.54	11.07
FSTL1	-1.33	-1.38	8.97	3.36
SLC4A7	-1.33	-1.37	5.42	2.04
GPSM1	-1.29	-1.37	5.38	2.03
CACNB3	-1.19	-1.37	0.81	0.30
IKZF2	-1.31	-1.36	2.13	0.81
VASH2	-1.30	-1.36	5.04	1.91
ST3GAL6	-1.26	-1.36	2.07	0.79
TCEA3	-1.23	-1.36	2.83	1.08
HMGA2	-1.32	-1.35	12.61	4.80
ARHGEF12	-1.31	-1.35	4.50	1.71

RFTN1	-1.27	-1.35	3.80	1.45
BICD1	-1.20	-1.35	1.01	0.39
UBASH3B	-1.31	-1.35	6.62	2.53
LAMA5	-1.27	-1.34	1.09	0.42
TCF7	-1.20	-1.34	1.18	0.45
HS2ST1	-1.31	-1.34	10.90	4.19
RAD23A	-1.29	-1.34	14.87	5.72
PTK7	-1.29	-1.34	7.79	3.01
TRPV4	-1.19	-1.33	1.06	0.41
NCKAP1	-1.26	-1.33	2.85	1.10
MIR155HG	-1.22	-1.33	3.89	1.51
AMOT	-1.20	-1.32	0.72	0.28
PDE4B	-1.25	-1.32	3.49	1.36
TAL1	-1.26	-1.32	3.94	1.54
FAM171B	-1.21	-1.31	1.88	0.74
PAQR7	-1.25	-1.31	7.31	2.88
PXN	-1.27	-1.31	11.48	4.52
DHRS12	-1.15	-1.31	2.12	0.84
RGPD4	-1.28	-1.30	24.16	9.56
RNF130	-1.27	-1.30	47.37	18.77
MAST4	-1.23	-1.29	1.49	0.59
ST3GAL5	-1.22	-1.29	5.90	2.36
FOS	-1.19	-1.28	3.74	1.50
AIF1L	-1.25	-1.28	21.68	8.68
METTL9	-1.25	-1.28	15.87	6.36
PADI4	-1.15	-1.27	1.89	0.76
UBE2E2	-1.20	-1.27	8.43	3.40
REEP2	-1.10	-1.27	1.08	0.44
IQGAP2	-1.22	-1.27	4.74	1.92
PAQR6	-1.14	-1.27	2.77	1.12
KBTBD8	-1.19	-1.27	2.40	0.97
BCAR1	-1.18	-1.26	3.17	1.29
FKBP5	-1.23	-1.26	21.84	8.87
BNIP3L	-1.22	-1.26	14.72	5.98
IRAK3	-1.24	-1.26	15.00	6.10
ZNF254	-1.21	-1.25	7.06	2.88
SH3BP5	-1.12	-1.25	1.12	0.46
LBR	-1.24	-1.25	76.15	31.14
PLK3	-1.07	-1.25	0.86	0.35

SLC39A10	-1.20	-1.24	6.54	2.69
SETD7	-1.20	-1.24	5.48	2.25
CASK	-1.13	-1.23	0.74	0.30
SH3RF2	-1.06	-1.23	0.73	0.30
PLEKHG2	-1.19	-1.23	4.35	1.81
ELK3	-1.14	-1.23	3.84	1.59
DSCC1	-1.17	-1.22	10.07	4.19
STOX1	-1.14	-1.22	2.90	1.21
RRM1	-1.19	-1.22	34.62	14.50
PITX1	-1.17	-1.21	13.61	5.72
TOX	-1.09	-1.21	1.11	0.46
TRIM24	-1.19	-1.21	42.99	18.10
DIXDC1	-1.14	-1.21	2.41	1.02
ARL2BP	-1.16	-1.20	13.21	5.58
IFT81	-1.09	-1.20	2.06	0.88
SLC26A2	-1.17	-1.19	15.61	6.65
SRP9	-1.17	-1.19	128.00	54.67
LHFP	-1.03	-1.18	1.22	0.52
PECAM1	-1.14	-1.18	8.48	3.63
CLDN10	-1.10	-1.18	3.88	1.66
CD44	-1.16	-1.18	30.46	13.07
TEAD2	-1.05	-1.18	1.81	0.78
SMAD6	-1.06	-1.18	1.56	0.67
PALM	-1.13	-1.18	10.46	4.50
RASL10B	-1.10	-1.18	3.41	1.47
TTC7B	-1.13	-1.17	11.60	5.02
DUSP9	-1.03	-1.17	1.38	0.60
SMYD3	-1.11	-1.17	14.01	6.08
CTXN1	-1.08	-1.16	7.19	3.12
SNN	-1.08	-1.16	2.45	1.06
IPO11-LRRC70	-1.11	-1.16	13.62	5.93
SPHAR	-1.07	-1.16	7.41	3.22
CEP85L	-1.10	-1.16	2.05	0.89
CDR2L	-1.06	-1.16	1.61	0.70
LRP12	-1.07	-1.16	1.72	0.75
DNAJB5	-1.09	-1.16	6.58	2.88
FMNL3	-1.09	-1.15	1.50	0.66
C9orf131	-1.04	-1.15	1.69	0.74
AKAP2	-1.09	-1.15	2.74	1.20

PALM2-AKAP2	-1.09	-1.15	2.49	1.09
PSD3	-1.08	-1.15	1.38	0.61
TMIGD2	-1.11	-1.15	24.95	10.94
FAM35A	-1.12	-1.15	17.67	7.75
YES1	-1.12	-1.15	12.91	5.67
BRSK1	-1.02	-1.15	1.15	0.50
ADAM22	-1.08	-1.15	1.32	0.58
SLC16A10	-1.07	-1.14	3.68	1.62
LHFPL2	-1.01	-1.13	0.79	0.35
FAM95B1	-1.03	-1.13	3.90	1.73
FSCN1	-1.11	-1.13	50.35	22.36
HDGFRP3	-1.08	-1.13	10.31	4.58
TMEM87B	-1.08	-1.13	4.24	1.88
OTUB2	-1.06	-1.13	3.04	1.35
FYTTD1	-1.10	-1.13	20.82	9.27
RNF157	-1.09	-1.13	7.57	3.37
RGPD5	-1.11	-1.12	38.58	17.24
RGPD6	-1.11	-1.12	36.41	16.29
SLC35G1	-1.02	-1.12	2.49	1.12
RUNX1T1	-1.10	-1.12	18.80	8.41
RGPD8	-1.11	-1.12	37.74	16.90
GCNT2	-1.03	-1.12	1.61	0.72
FAM35BP	-1.07	-1.12	6.70	3.00
SDCCAG8	-1.09	-1.12	25.48	11.45
TMEM200A	-1.02	-1.10	2.82	1.28
SLC38A1	-1.09	-1.10	24.80	11.23
GBP4	-1.06	-1.10	4.84	2.20
TBC1D4	-1.05	-1.10	3.82	1.73
ZNF138	-1.02	-1.10	3.70	1.68
PRKCH	-1.01	-1.10	2.09	0.95
FZD6	-1.06	-1.10	11.59	5.28
MINPP1	-1.04	-1.09	10.40	4.74
CCNE2	-1.04	-1.09	7.36	3.36
SLC37A3	-1.03	-1.09	5.68	2.60
GCNT1	-1.05	-1.09	7.14	3.27
ABCC4	-1.05	-1.09	9.54	4.37
TET1	-1.05	-1.08	4.44	2.04
CALM2	-1.06	-1.08	172.72	79.80
TMED5	-1.05	-1.07	14.13	6.53

HOMER1	-1.00	-1.07	2.88	1.33
FAM35DP	-1.02	-1.07	7.42	3.44
UBE2K	-1.05	-1.07	21.10	9.79
ADCY6	-1.00	-1.07	2.26	1.05
SLC25A13	-1.01	-1.06	7.99	3.72
DRAM1	-1.05	-1.06	51.85	24.16
SLC5A3	-1.03	-1.06	6.37	2.97
ATP8B2	-1.02	-1.06	6.84	3.20
STMN1	-1.03	-1.05	122.21	57.51
HK1	-1.02	-1.05	28.69	13.53
UNG	-1.01	-1.03	41.41	19.70
MAPK6	-1.00	-1.03	17.30	8.24
GNPTAB	-1.01	-1.02	59.18	28.35

GFOLD (generalized fold change), a reliable statistics for expresson changes based on the posterior distribution of log fold change given by the tool GFOLD.

致谢

感谢王侃侃老师在硕士研究生阶段对我的教育和指导。王老师带我从一个对生物信息毫不了解的药学本科生走进了生物信息学的领域，并让我热爱上了这一领域。王老师在课题进行的过程中，给了非常多并且有益处的指导，和我们一起探讨课题、研究文章。最后在论文的整理和发表过程中，王老师一丝不苟和追求细节与完美的精神和品质使我终生收益。

感谢课题的合作者李易真师姐，在 AML1/ETO 的这个课题上，我们两个互相讨论、互相学习。虽然自己不做实验，但是李易真师姐给我讲了很多实验的原理以及方法，让我能更好的完成这一课题。

感谢谭云师兄，从大学一年级到现在，作为我的学长和师兄一直关心帮助着我。尤其感谢能够推荐我来这里，让我有机会能开启我的研究生生涯。

感谢杨文涛师兄，最开始的生物信息学入门多亏了杨文涛师兄的指点。感激之情难以言表。

感谢实验室照顾过我的师兄师姐，张辉师兄、金雯师姐、王业伟师兄、赵旭杰师兄、杨贤雯师姐、刘静秋师姐、史冬师兄、王朝师兄、邓望龙师兄、王果师姐、陈明杰师兄、许海峰师兄、张善镇师兄。

感谢实验室的各位师弟师妹们，韦舒勇、王晓玲、马雪菲、马云林、陈儒、张荣胜、李精明，我们共同成长。

感谢这三年来在生活上一直照顾和陪伴我的各位老师和同学。我的室友林如海，对我的生活诸多的照顾；我的同学陈志未、杨舟、晏伟伟、戴钰俊、熊伟昕等；我打篮球的好伙伴盖东征、王鹏然、李咸洋、谢银银、张元亮、郭河洲、张翔、李东河等；我打羽毛球的球友施静艺老师、郭淑杰老师、刘炳亚老师、范立权老师、张琳老师、黄金艳老师、顾朝晖师兄、张金丽师姐、王伯言师姐、张自冠等；我踢足球的各位朋友杨林森、刘丹、邱青松等。

感谢实验室范惠咏老师的照顾。感谢科教处徐勤毅老师、朱佳敏老师的帮助。

在研究生期间，我收获了爱情，感谢张然然的陪伴，有了你让我了解到工作或者科研并不是我生活的全部。感谢我的父母对我无私地支持，想到你们永远是我前进最大的动力。

最后，再次谢谢大家，我爱你们。

硕士就读期间发表论文

1. Hui Zhang, Jianqing Mi, Hai Fang, Zhao Wang, Chun Wang, Lin Wu, Bin Zhang, Mark Minden, Wentao Yang, **Huanwei Wang**, Junmin Li, Xiaodong Xi, Saijuan Chen, Ji Zhang, Zhu Chen, and Kankan Wang. Preferential eradication of acute myelogenous leukemia stem cells by fenretinide. *Proc. Natl. Acad. Sci. U. S. A.*, 2013; 110: 5606-5611.
2. Yizhen Li, **Huanwei Wang**, Xiaoling Wang, Wen Jin, Yun Tan, Hai Fang, Saijuan Chen, Zhu Chen, and Kankan Wang. Genome-wide studies identify a novel interplay between AML1 and AML1/ETO in t(8;21) acute myeloid leukemia. *Blood*, 2015 (submitted).