(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116343917 B
(45)授权公告日 2023.11.10

(21)申请号 202310326955.5

(22)申请日 2023.03.22

(65)同一申请的已公布的文献号
申请公布号 CN 116343917 A

(43)申请公布日 2023.06.27

(73)专利权人 电子科技大学长三角研究院(衢州)
地址 324003 浙江省衢州市柯城区芹江东路288号1幢18楼

(72)发明人 刘利 韩璐 邹权 丁漪杰

(74)专利代理机构 浙江永鼎律师事务所 33233
专利代理师 陆永强

(51)Int.Cl.
G16B 20/30(2019.01)

(56)对比文件
CN 101918578 A,2010.12.15
US 2019085324 A1,2019.03.21
US 2020190494 A1,2020.06.18
WO 2022147296 A1,2022.07.07
CN 107368701 A,2017.11.21
US 2018016314 A1,2018.01.18

审查员 柳倩

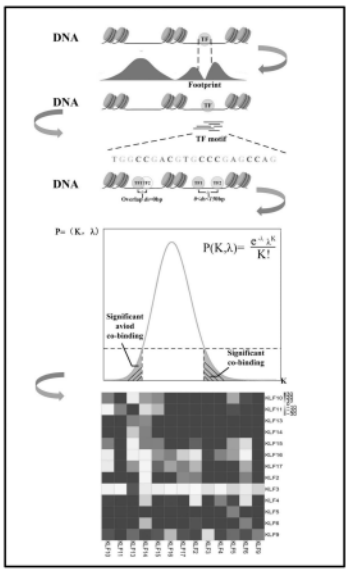
权利要求书2页 说明书7页 附图11页

(54)发明名称

一种基于ATAC-seq足迹识别转录因子共定位的方法

(57)摘要

本方案公开了一种基于ATAC-seq足迹识别转录因子共定位的方法,首先根据ATAC-seq数据,使用HINT-ATAC方法进行数据预处理,然后基于泊松分布构建识别转录因子共定位模型,基于距离ds实现协同结合和竞争结合的转录因子对的识别,通过仅使用ATAC-seq数据作为输入,就能识别协同结合和竞争结合的转录因子对,为进一步应用于多个细胞系提供基础。



CN 116343917 B

1.一种基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,该方法包括:

S1.收集并下载待识别目标的ATAC-seq数据,以获取原始的染色质开放性测序数据;

S2.基于步骤S1获取的数据文件,使用足迹分析工具分析转录因子足迹在待识别目标基因组中的坐标数据;

S3.利用转录因子数据库的转录因子motif与足迹坐标数据进行匹配,以获取每个结合位点的具体转录因子种类;

所述的转录因子数据库收录有不同生物的转录因子与结合位点及结合方式;

S4.通过距离计算工具计算每两种转录因子之间的距离ds;

根据计算的ds,以第一距离阈值为准确定两种转录因子的共定位数k1;

根据计算的ds,以第二距离阈值为准确定两种转录因子的共定位数k2;

S5.基于泊松分布构建识别转录因子共定位模型;

使用识别转录因子共定位模型分别计算k1、k2两种情况下的概率值P;

根据概率值P和阈值P'分别筛选两种情况下具有显著性的转录因子对;

S6.对于第二距离阈值的情况,将k1大于期望值的显著性转录因子对判断为协同结合的转录因子对;

对于第一距离阈值情况,将k2大于期望值的显著性转录因子对判断为竞争结合的转录因子对;

根据竞争结合的转录因子对判断各协同结合的转录因子对是否同时属于竞争结合,若是,则将相应转录因子对判断为既竞争又协同。

2.根据权利要求1所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,步骤S5中,显著性的转录因子对包括显著共定位的转录因子对和显著拒绝共定位的转录因子对。

3.根据权利要求2所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,基于泊松分布构建识别转录因子共定位模型为:

$$P(k;n,m,N)=\frac{e^{-\lambda}\lambda^{k}}{k!} \quad (1)$$

$$\lambda=\frac{nm}{N} \quad (2)$$

其中公式(1)中k为两转录因子在阈值范围内的共定位数,n,m分别为两种转录因子各自的定位数目,N代表待识别目标总共的结合位点,λ为期望值;

通过公式(1)分别计算在两种距离阈值下,每两种转录因子对被分别判断为共定位或拒绝共定位的概率;

基于概率值分别筛选出两种距离阈值下具有显著性特征的转录因子对。

4.根据权利要求3所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,每两种转录因子,使用相距最近的转录因子间距离作为该两种转录因子间的距离。

5.根据权利要求4所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,在每种阈值情况下,根据每种转录因子匹配有几个坐标确定该转录因子的定位数目。

6.根据权利要求5所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,设定的两种距离阈值,第一距离阈值为ds=0,第二距离阈值为0<ds<150。

7.根据权利要求2所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,步骤S1中,所述的待识别目标为目标细胞系或目标组织。

8.根据权利要求2所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,步骤S2中,所使用足迹分析工具为HINT-ATAC工具,且步骤S1获取的是与HINT-ATAC工具兼容的ATAC-seq narrowpeak格式文件。

9.根据权利要求8所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,步骤S3中,使用HINT-ATAC软件的motif-analysis模块实现所述的匹配工作。

10.根据权利要求9所述的基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,步骤S3中,所述的转录因子数据库采用JASPAR数据库。

一种基于ATAC-seq足迹识别转录因子共定位的方法

技术领域

[0001] 本方案属于计算机生物学技术领域,提出了一种基于ATAC-seq足迹识别转录因子共定位的方法。

背景技术

[0002] 转录因子(transcription factor,TF)是一类序列特异性DNA结合蛋白,能够结合在靶基因上游的转录因子结合位点序列,参与调控基因转录过程,从而保证目的基因以特定的强度在特定的时间与空间表达。一般来说,转录因子以组合的形式调控高等生物基因的表达,大多数的转录因子必须共同发挥作用才能完成转录任务。因此,在基因研究以及基因病研究中,获得显著的共定位转录因子对就显得很有必要。现有的转录因子共定位识别方法包括基于转录因子ChIP-seq数据或者motif匹配的统计检验方法。

[0003] 基于ChIP-seq的方法,首先收集细胞系所有转录因子的ChIP-seq实验数据,确定各转录因子在全基因组的结合位点,对两两转录因子结合位点的关系进行统计检验,获得显著的共定位转录因子对。

[0004] 转录因子motif匹配方法一般是利用ATAC-seq测序数据得到染色质开放性区域,在这些区域上实施转录因子motif扫描,识别潜在的转录因子结合位点,进而通过统计分析识别共定位转录因子对。

[0005] 基于ChIP-seq的方法所需的输入数据量大,即,一种细胞系如果想做多个转录因子,就要有多个转录因子ChIP-seq实验数据作为输入,对所研究的细胞系或组织需要几百上千的转录因子ChIP-seq实验数据。目前,如此多的实验数据仅能在有限的细胞系中获得,所以存在实验数据采集的局限性。可变地,基于motif匹配的方法仅利用一种实验数据ATAC-seq即能分析转录因子的共定位,但是无法对共定位定义区分重叠定位还是近邻定位,从而均不能反映转录因子对是竞争结合还是协同结合,而区分转录因子是竞争关系还是协作关系对理解基因转录调控的分子机制是至关重要的。

发明内容

[0006] 本方案的目的是针对上述问题,提供一种基于ATAC-seq足迹识别转录因子共定位的方法,首先根据ATAC-seq数据,使用HINT-ATAC方法进行数据预处理,然后基于泊松分布构建识别转录因子共定位模型,基于距离ds实现协同结合和竞争结合的转录因子对的识别,通过仅使用ATAC-seq数据作为输入,就能识别协同结合和竞争结合的转录因子对,为进一步应用于多个细胞系提供基础。

[0007] 一种基于ATAC-seq足迹识别转录因子共定位的方法,其特征在于,该方法包括:

[0008] S1.收集并下载待识别目标的ATAC-seq数据,以获取原始的染色质开放性测序数据;一个细胞系或组织有一个对应的ATAC-seq数据,也就是说,对于一个细胞系,只需要一种ATAC-seq数据。

[0009] S2.基于步骤S1获取的数据文件,使用足迹分析工具分析转录因子足迹在待识别

目标基因组中的坐标数据;ATAC-seq数据是开放区数据,通过footprint分析能识别到基因组中哪些位点是转录因子的结合位点,但是不确定是哪种录因子。

[0010] S3.利用转录因子数据库的转录因子motif与足迹坐标数据进行匹配,以获取每个结合位点的具体转录因子种类;通过footprint+motif能够得到某种转录因子的结合位点。

[0011] 所述的转录因子数据库收录有不同生物的转录因子与结合位点及结合方式;

[0012] S4.通过距离计算工具计算每两种转录因子之间的距离ds;

[0013] 根据计算的ds,以第一距离阈值为准确定两种转录因子的共定位数k1;

[0014] 根据计算的ds,以第二距离阈值为准确定两种转录因子的共定位数k2;

[0015] S5.基于泊松分布构建识别转录因子共定位模型;

[0016] 使用识别转录因子共定位模型分别计算k1、k2两种情况下的概率值P;

[0017] 根据概率值P和阈值P'分别筛选两种情况下具有显著性的转录因子对;

[0018] S6.对于第二距离阈值的情况,将k1大于期望值的显著性转录因子对判断为协同结合的转录因子对;

[0019] 对于第一距离阈值情况,将k2大于期望值的显著性转录因子对判断为竞争结合的转录因子对;

[0020] 根据竞争结合的转录因子对判断各协同结合的转录因子对是否同时属于竞争结合,若是,则将相应转录因子对判断为既竞争又协同。

[0021] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,步骤S5中,显著性的转录因子对包括显著共定位的转录因子对和显著拒绝共定位的转录因子对。

[0022] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,基于泊松分布构建识别转录因子共定位模型为:

[0023] $$P(k;n,m,N)=\frac{e^{-\lambda}\lambda^{k}}{k!} \quad (1)$$

[0024] $$\lambda=\frac{nm}{N} \quad (2)$$

[0025] 其中公式(1)中k为两转录因子在阈值范围内的共定位数,n,m分别为两种转录因子各自的定位数目,N代表待识别目标总共的结合位点,λ为期望值;

[0026] 通过公式(1)分别计算在两种距离阈值下,每两种转录因子对被判断为共定位或拒绝共定位的概率;

[0027] 基于概率值分别筛选出两种距离阈值下具有显著性特征的转录因子对。

[0028] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,每两种转录因子,使用相距最近的转录因子间距离作为该两种转录因子间的距离。即,在一个细胞系中,一种转录因子可能有多个,那么每两种转录因子之间就可能有多个距离,这里取最短的距离作为两种转录因子间的距离。

[0029] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,根据每种转录因子匹配有几个坐标确定该转录因子的定位数目。

[0030] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,设定的两种距离阈值,第一距离阈值为ds=0,第二距离阈值为0<ds<150。

[0031] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中,步骤S1中,所述的待

识别目标为目标细胞系或目标组织。

[0032] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中，步骤S2中，所使用足迹分析工具为HINT-ATAC工具，且步骤S1获取的是与HINT-ATAC工具兼容的ATAC-seq narrowpeak格式文件。

[0033] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中，步骤S3中，使用HINT-ATAC软件的motif-analysis模块实现所述的匹配工作。

[0034] 在上述的基于ATAC-seq足迹识别转录因子共定位的方法中，步骤S3中，所述的转录因子数据库采用JASPAR数据库。

[0035] 本方案的优点在于：

[0036] (1)本发明提出了一种基于ATAC-seq足迹识别转录因子共定位的方法，该方法仅需使用ATAC-seq数据作为输入就能够实现对转录因子共定位的识别，可以应用于多个细胞系，为进一步探索转录因子以组合的模式与DNA相互作用提供方法支持；

[0037] (2)本发明基于泊松分布背景模型，计算共定位P值矩阵，识别具有统计显著性的转录因子共定位，从统计学角度排除随机背景的影响，有效识别转录因子共定位；

[0038] (3)本发明利用footprint高分辨率数据，有效提高了转录因子结合位点识别的准确性；

[0039] (4)本发明通过设定两个阈值，在泊松分布筛选的显著性转录因子对的基础上，对其实现了重叠定位和近邻定位的区分，不仅能够实现共定位转录因子对的识别，而且还能够区分其属于竞争结合还是协同结合，对于转录因子调控的分子机制研究有着重要的意义。

附图说明

[0040] 图1为本发明实施例提供的基于ATAC-seq足迹识别转录因子共定位的方法流程图；

[0041] 图2为本发明实施例提供的当距离阈值为ds＝0bp时转录因子聚类生成的P值矩阵热图；

[0042] 图3为本发明实施例提供的当距离阈值为0<ds<150bp时转录因子聚类生成的P值矩阵热图；

[0043] 图4A为本发明实施例提供的距离阈值为ds＝0bp时，按转录因子名字的字母顺序排列的P值矩阵热图；

[0044] 图4B为本发明实施例提供的距离阈值为0bp<ds<150bp时，按转录因子名字的字母顺序排列的P值矩阵热图；

[0045] 图4C为图4A中FOS_JUN家族转录因子共定位的P值矩阵热图；

[0046] 图4D为图4B中FOS_JUN家族转录因子共定位的P值矩阵热图；

[0047] 图4E为FOS_JUN家族motif位点信息含量logo图；

[0048] 图5A为本发明实施例提供的距离阈值为ds＝0bp时，KLF家族转录因子共定位P值矩阵热图；

[0049] 图5B为本发明实施例提供的0bp<ds<150bp时，KLF家族转录因子共定位P值矩阵热图；

[0050] 图6A-图6D为分别为ChIP-seq、ChIP-exo、footprint、ATAC-seq四种数据峰的长度分布对比；

[0051] 图7A为ChIP-seq+motif与ChIP-exo比较的维恩图；

[0052] 图7B为footprint+motif与ChIP-exo比较的维恩图。

具体实施方式

[0053] 下面结合附图和具体实施方式对本方案做进一步详细的说明。

[0054] 本实施例给出了一种基于ATAC-seq足迹识别转录因子共定位的方法，如图1所示，具体包含以下步骤：

[0055] S1.收集并下载待识别目标的ATAC-seq数据，如K562细胞系，以获取原始的染色质开放性测序数据。ENCODE包含370个组织或细胞系的ATAC-seq测序数据可用，本实施例从ENCODE获取K562细胞系的ATAC-seq数据。

[0056] S2.根据下载的ATAC-seq narrowpeak格式文件，使用HINT-ATAC工具得到转录因子footprint(足迹)在基因组中的坐标数据。

[0057] HINT-ATAC为LINUX系统下的软件RGT，RGT是一个开源库，HINT-ATAC是RGT库中的一个开源软件，HINT-ATAC可以用于进行footprint分析，通过footprint分析得到footprint在基因组中的坐标数据，可用于获取转录因子在全基因组上结合情况，其具体分析方式直接采用现有技术即可，不在此赘述。

[0058] 此时能够识别到基因组中哪些位点是转录因子的结合位点，但不确定是哪些转录因子。

[0059] S3.利用JASPAR数据库的转录因子motif在上述的footprint位置进行匹配。具体为，对上述footprint坐标数据与JASPAR数据库匹配转录因子motif，分辨每一个结合位点具体的转录因子种类，具体方法是使用HINT-ATAC软件的motif-analysis模块进行匹配，阈值可选0.0001。

[0060] S4.将匹配到的N种转录因子根据转录因子名字进行排列，以快速区分转录因子家族；针对K562细胞系，总共匹配了633种转录因子；

[0061] S5.motif匹配完以后，可以得到每两种转录因子在足迹区域匹配的次数(m,n)，也即每种转录因子自己的个数(m,n)；

[0062] 随后，将footprint匹配的各转录因子motif坐标数据作为输入，利用bedtools计算得到两种转录因子共定位的个数，记为k值。

[0063] 每种转录因子匹配有几个坐标就有几次匹配次数，可以依此得到每种转录因子的匹配次数，即对于每两种转录因子，可以得到上述匹配次数m、n。

[0064] 计算每两种转录因子的共定位数k具体包括以下步骤：

[0065] S5-1.对两种转录因子，利用bedtools closest-d得到相距最近的转录因子间距离ds；

[0066] 如TFA与TFB为任意两种转录因子，TFA为转录因子A的位点信息，有三列，分别为染色体，起始位点，终止位点，B也是一样的。两个位点文件经过bedtools closest处理后，找到与TFA最近的TFB的转录因子，生成新的文件，文件为7列，前三列为TFA位置，后三列为TFB，第7列为最近的距离，即ds。

[0067] S5-2.将ds取值分为ds＝0和0<ds<150两种情况，设定这两种阈值，分别计算k值；ds＝0表示两个转录因子位点有重叠。

[0068] ds＝0和0<ds<150是按多数开放区域的长度评估所得的经验阈值，这里将ds＝0作为第一距离阈值，0<ds<150作为第二距离阈值。在实际应用时，本领域技术人员也可以将其改为其它数值分别作为第一距离阈值和第二距离阈值。

[0069] 根据两种距离阈值，两种转录因子共定位的个数可能会发生改变，故每两种转录因子将得到两个k值，一个对应第一距离阈值，一个对应第二距离阈值。

[0070] S5-3.根据633种转录因子两两配对计算，形成两个633×633的k值矩阵；k是共定位符合距离阈值的个数，可以根据k与期望值λ的比较，判定是否共定位；

[0071] S6.通过上述m,n,k值，基于泊松分布，得到两两转录因子共定位的显著性P值矩阵，根据P值分布判定是否显著共定位。也就是说，这里会得到两种情况下的显著性P值矩阵，一个是距离阈值为ds＝0的情况，一个是距离阈值为0<ds<150的情况。

[0072] 其中泊松分布，其计算方法为：

[0073] $$P(k;n,m,N)=\frac{e^{-\lambda}\lambda^{k}}{k!} \quad (1)$$

[0074] $$\lambda=\frac{nm}{N} \quad (2)$$

[0075] 其中公式(1)中k为两转录因子在距离阈值范围内共定位数，n,m分别为两种转录因子各自的定位数目，N代表的是总共的结合位点，表示整个基因组中可用于转录因子(TF)结合的区域的数量。N是根据ATAC-seq和footprint两种数据共同判定的，如本例中，K562细胞系的633种转录因子在269997个footprint区域有匹配，则N为269997。λ为期望值，根据n、m和N得到。

[0076] 在双尾泊松分布中，低P值表明两转录因子在基因组中定位非随机，可以代表两种显著性，一种是显著的共定位，另一种则是显著的拒绝共定位，如图1中左边阴影部分为显著的拒绝共定位，右边阴影部分为显著的共定位，区分显著或不显著的阈值P’值由本领域技术人员根据经验确定，如可以取0.01，也可以根据总数，取总数的一定比例，如取总数的1％。通过该方式筛选出具有显著性的转录因子对，筛选出的转录因子对可能是显著的共定位，也可能是显著的拒绝共定位。

[0077] 然后使用k与期望值λ的关系对筛选出的转录因子对进行进一步判断，如果k高于期望值，认为两转录基因(TF)倾向于在基因组上共定位，如果k低于或等于期望值，则认为两TF倾向于在基因组上拒绝共定位。通过以上方式，针对距离阈值分别为ds＝0和0<ds<150将分别得到两种结果，每种结果包含若干对被判断为共定位的转录因子对和若干对被判断为拒绝共定位的转录因子对。

[0078] S7.筛选阈值为0<ds<150时判断为共定位的转录因子对，得到协同结合的转录因子对；

[0079] 筛选阈值ds＝0时判断为共定位的转录因子对，得到竞争结合的转录因子对；

[0080] 根据竞争结合的转录因子对判断各协同结合的转录因子对是否同时属于竞争结合，若是，则将相应转录因子对判断为既竞争又协同，并将其从竞争结合的转录因子对和协同结合的转录因此对中剔除。

[0081] 如此，便筛选出了协同结合的转录因子对，竞争结合的转录因子对和既竞争结合又协同结合的转录因子对。也就是说，本方案提出使用分辨率更高的footprint+motif方法，同时结合基于统计学的泊松分布通过设定两个阈值的方式，不仅能够更准确地筛选出共定位的转录因子对，而且还能够对共定位的转录因子对区分其是竞争结合还是协同结合，可用于帮助理解基因转录调控的分子机制，协助当前细胞系和组织的基因研究。

[0082] 图2所示为本实施例提供的当距离阈值ds＝0bp时转录因子聚类生成的P值矩阵热图，右上角的+30表示显著的共定位，-30表示显著的拒绝共定位。原图为彩色显示，显著的共定位为蓝色，显著的拒绝共定位为红色，越不显著，颜色越浅。蓝色主要集中在对角线处，其它位置有部分偏蓝色，表示共定位的TF对主要集中对角线处，但是其他位置也有。在灰度处理以后，图2中明显黑色的一簇簇表示显著的共定位，深灰色表示显著的拒绝共定位，白色和浅灰色是没有那么显著的TF对。可以看到，对角线显示本身与本身的竞争结合(同一转录因子竞争同一位点)，对角线上聚类簇大部分为同一基因家族，其具有相同或相似的motif。

[0083] 图3所示为本实施例提供的当0<ds<150bp转录因子聚类生成的P值矩阵热图，同样右上角的+30表示显著的共定位，-30表示显著的拒绝共定位。原图为彩色显示，显著的共定位为蓝色，显著的拒绝共定位为红色，显示了协同结合的转录因子对。在该图中，大部分为偏蓝色，少量偏红色，表示大部分TF对表现为共定位，灰度处理以后，由于参与的转录因子有633种之多，即横坐标633个，纵坐标633个，数据量巨大，已经很难看到区别，为此，为了使读者更好地理解，本方案将在下面以FOS家族和KLF家族为例进行详细说明。

[0084] 图4A和图4B分别为本实施例提供的距离阈值为ds＝0bp和0bp<ds<150bp时，按转录因子名字的字母顺序排列的P值矩阵热图。原图仍为彩色显示，显著的共定位为蓝色，显著的拒绝共定位为红色，图4A、图4B与图2和图3类似，区别仅在于是否进行聚类、是否按照字母顺序排列，由于转录因子数量巨大，灰度处理后，其实仍然不清楚，将在下面以FOS家族和KLF家族为例进行详细说明。

[0085] 图4C和图4D是FOS_JUN家族的一个例子，图4C是图4A中FOS_JUN家族转录因子共定位的P值矩阵热图，图4D是图4B中FOS_JUN家族转录因子共定位的P值矩阵热图。对于同一家族的转录因子，转录因子之间有相同或相近的motif。图4E为FOS_JUN家族motif位点信息含量logo图，可以看到显著共定位的转录因子对的家族motif相似。

[0086] 为了更清晰的表示，图4C中，圆圈表示偏蓝色，点表示偏红色，颜色越浅，相应的显著性越弱，未标注的黑色表示显著共定位的蓝色，白色表示接近0的没有显著性。图4D中，标注圈的框表示偏蓝色其余为偏红色或没有显著性。

[0087] 图4D是阈值为0<ds<150时的P值矩阵热图，颜色较深的标注圈的框表示显著的共定位，从图4D可以看到，这样的框其实不多，表示此家族只有少量的协同结合转录因子对，如FOSB::JUNB与FOS，FOSB::JUNB(var.2)与FOS，FOS::JUN(var.2)与FOSL1等。

[0088] 图4C是阈值为ds＝0时候的P值矩阵热图，颜色较深的标注点的框表示显著拒绝共定位，可以看到图4C不存在这样的框，即不存在显著拒绝共定位的转录因子对，未标注的黑色框和颜色较深的标注圈的框表示显著共定位，可以看到图4C中有较多这样的框，表示该家族有较多的竞争结合的转录因子对，如FOSB..JUN与FOSL2::JUN，FOSB..JUN与FOSL2::JUND等。

[0089] 再根据竞争结合的转录因子对和协同结合的转录因子对,可以找出既竞争又结合的转录因子对,从图4C和图4D可以看到,此家族不存在这样的转录因子对。

[0090] 图5A和图5B分别为本实施例提供的距离阈值为ds＝0bp和0bp<ds<150bp时,KLF家族转录因子共定位P值矩阵热图,图5A是图4A中KLF家族转录因子共定位的P值矩阵热图,图5B是图4B中KLF家族转录因子共定位的P值矩阵热图。图5A中全部框显示较深的蓝色,表示距离阈值为ds＝0时候这一块均显示显著的共定位,各转录因子对均被判断为竞争结合。图5B中,用点表示偏红色,即偏拒绝共定位,其余未标注的偏蓝色,即共定位。以此为例,图5B中,颜色较深的未标注框(即显著共定位的转录因子对),将被筛选为协同结合的转录因子对。此外,可以看到,此处被判断为协同结合的转录因子对,有些在阈值为ds＝0bp时也表现为显著的共定位,即同时被判断为竞争结合,这样的转录因子对将被判断为既竞争又协同的转录因子对,如KLF10和KLF9,KLF11和KLF9等,表明本方法可以区分竞争结合和协同结合。

[0091] 本方案采用了footprint+motif方式识别转录因子结合位点,现有技术可以实现ChIP-seq+motif方式识别转录因子结合位,ChIP-seq和ChIP-exo均是专门定位具体某种转录因子在基因组中的结合位点,只是后者分辨率更高,能更准确的定位,现有技术为了提高ChIP-seq分辨率,通常将ChIP-seq和motif结合,但是我们知道该方法虽然能够实现转录因子结合位点的识别,却存在数据量大的问题。

[0092] 本实施例以更高分辨率的ChIP-exo数据为金标准比较验证footprint+motif在识别转录因子结合位点方面与传统方法ChIP-seq+motif是否具有同等效力。

[0093] 如图6A-图6D所示为ChIP-seq、ChIP-exo、footprint、ATAC-seq四种数据峰的长度分布对比,图6A显示了ChIP-seq数据的长度分布,图6B显示了ChIP-exo数据的长度分布,图6C是footprint数据长度分布,图6D是ATAC-seq数据长度分布,由图可知,footprint具有最高的数据分辨率,在15bp左右,而ChIP-exo长度分布在50bp左右,ChIP-seq的长度分布在250bp左右,所以在分辨率上来说,本方案采用的方法是存在优势的。

[0094] 这里进一步使用相对ChIP-seq具有更高分辨率的ChIP-exo数据作为金标准,将footprint+motif与ChIP-seq+motif数据比较。对于ChIP-exo数据,这里使用bwa默认参数将原始读数与参考基因组进行比对。双端测序数据用samtools rmdup效果很差,所以用picard工具的MarkDuplicates功能删除带有'MarkDuplates Remove Duplates＝TRUE'选项的PCR重复项。用MACS2判别reads比对后在基因组中形成的峰和峰顶,P值为0.001。经过call peak处理的ChIP-exo数据文件为narrowpeak格式。

[0095] 图7A和图7B所示分别为ChIP-seq+motif、footprint+motif与ChIP-exo比较的维恩图。每种方法中重叠部分的比例用百分比数字标记,由图可知,ChIP-exo作为转录因子结合位点金标准时,ChIP-seq+motif与footprint+motif两者相当,说明本方案使用一种ATAC-seq数据的footprint+motif数据可以替代需要使用大量数据的ChIP-seq+motif识别潜在的转录因子结合位点。

[0096] 本文中所描述的具体实施例仅仅是对本方案精神作举例说明。本方案所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替代,但并不会偏离本方案的精神或者超越所附权利要求书所定义的范围。

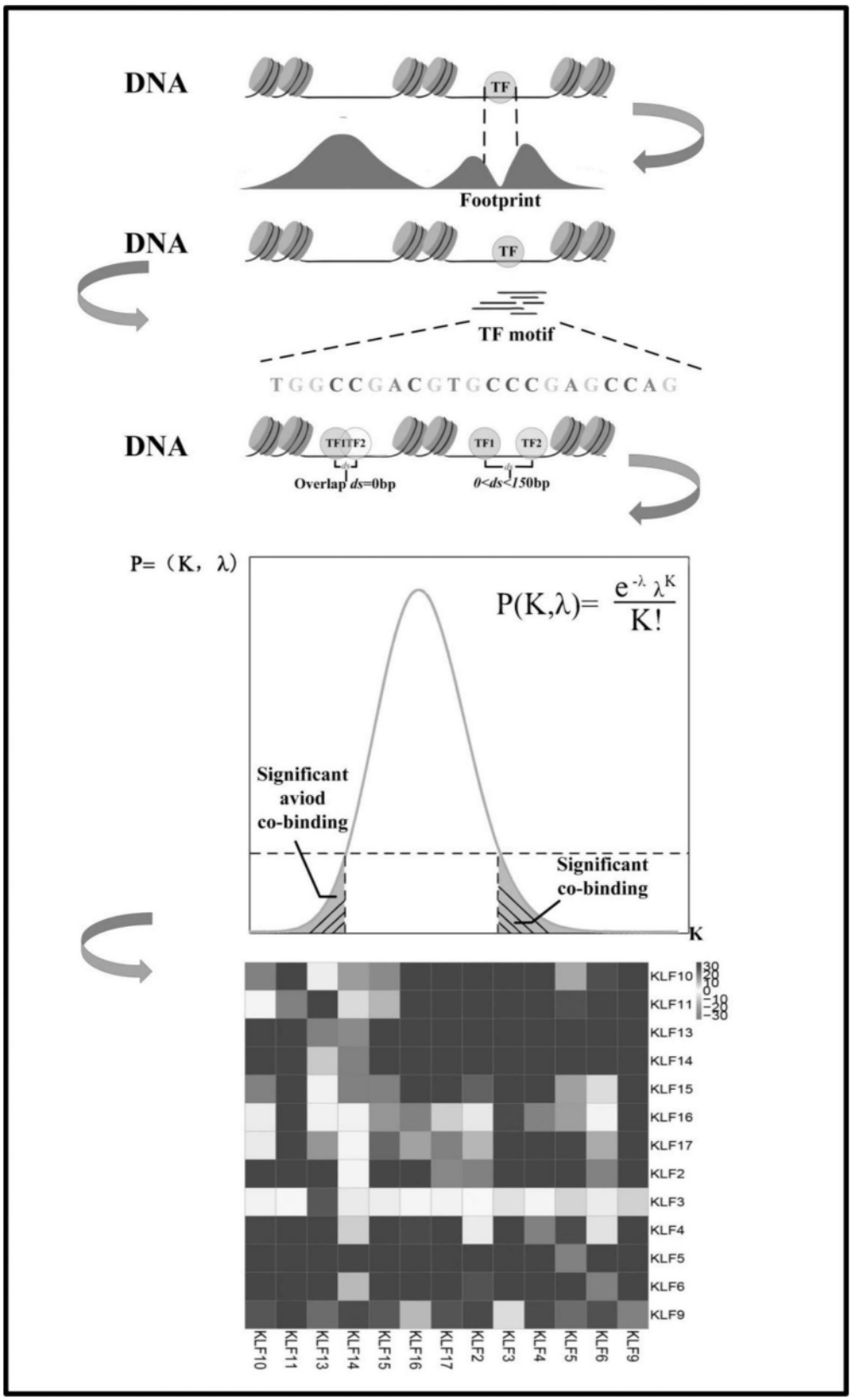

DNA
TF
Footprint
DNA
TF
TF motif
TGGCCGACGTGCCCGAGCCAG
DNA
TF1
TF2
TF1
TF2
Overlap ds=0bp
0<ds<150bp
P=（K，λ）
P(K,λ)= e^-λ λ^K / K!
Significant aviod co-binding
Significant co-binding
K
KLF10
KLF11
KLF13
KLF14
KLF15
KLF16
KLF17
KLF2
KLF3
KLF4
KLF5
KLF6
KLF9
30
20
10
0
−10
−20
−30

图1

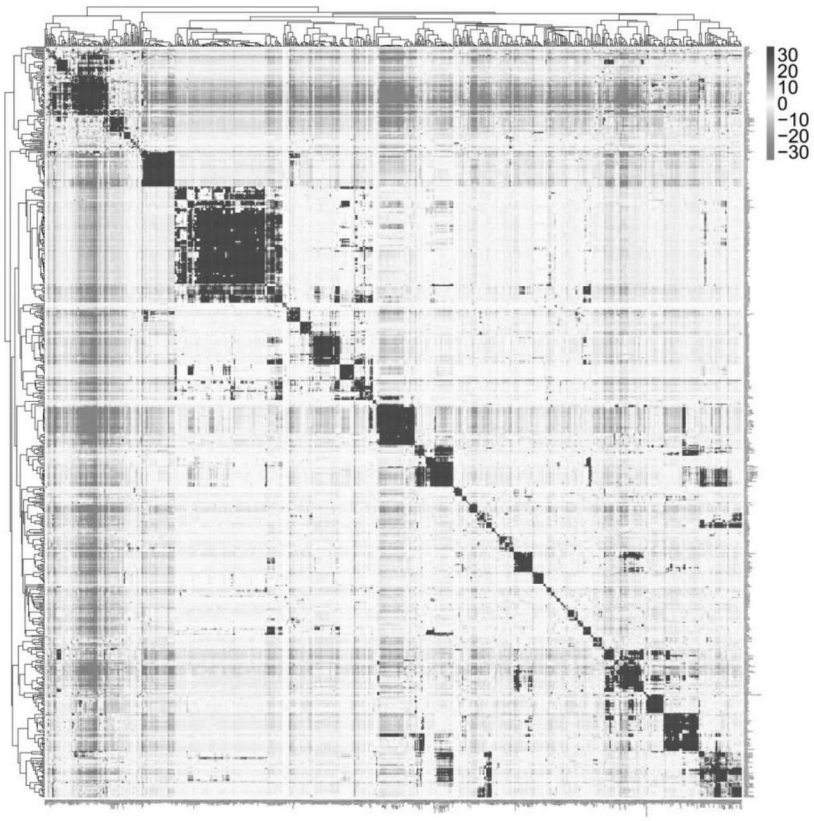


图2

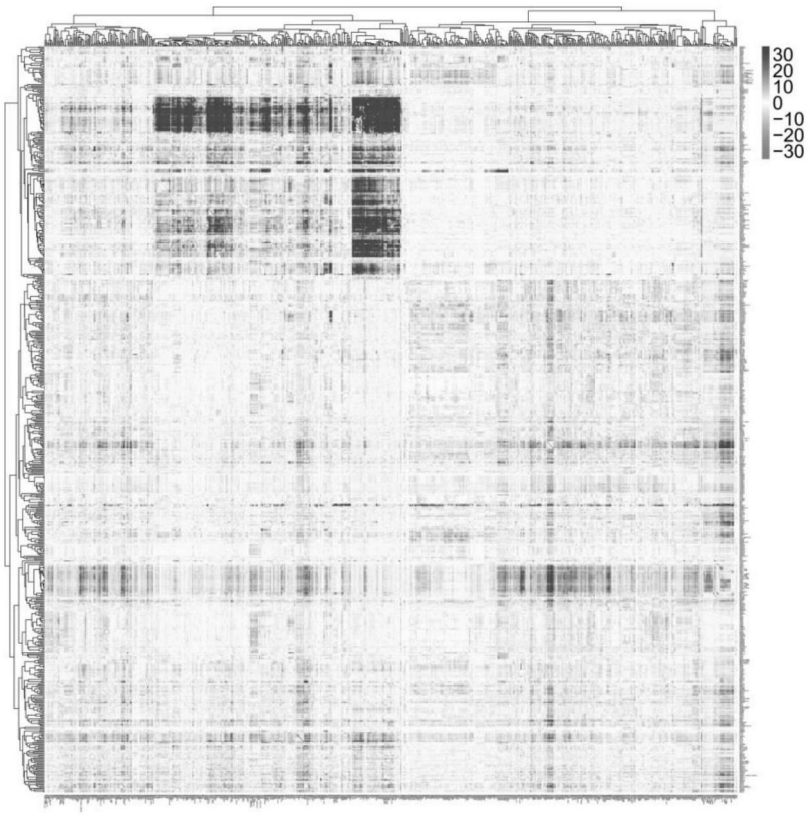


图3

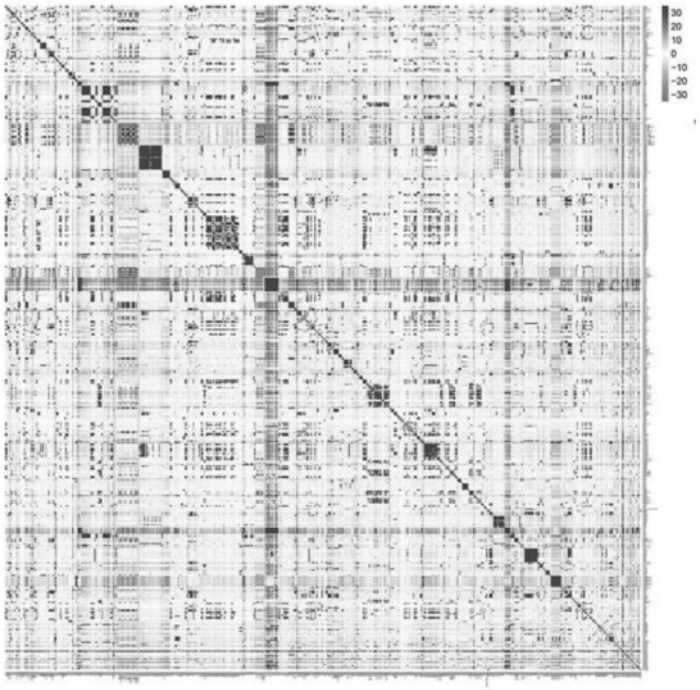


图4A

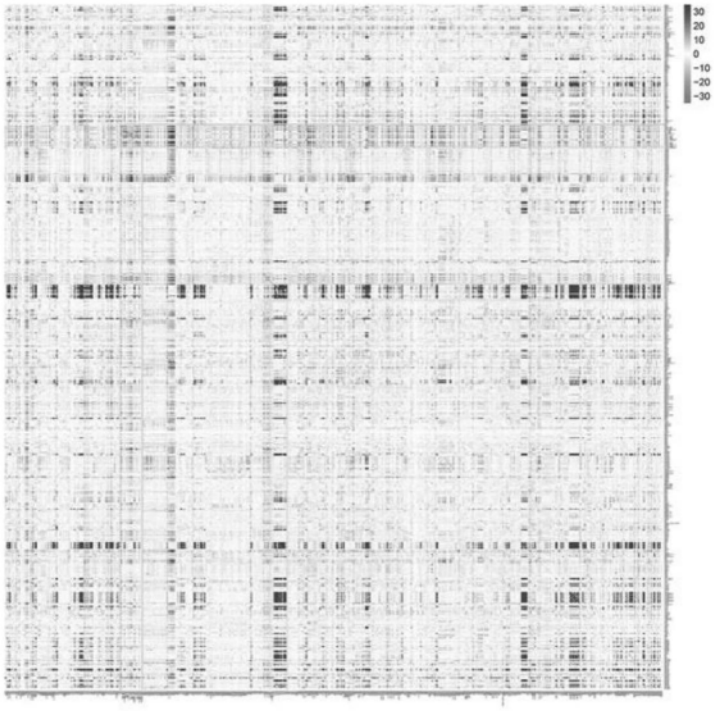

30
20
10
0
−10
−20
−30

图4B

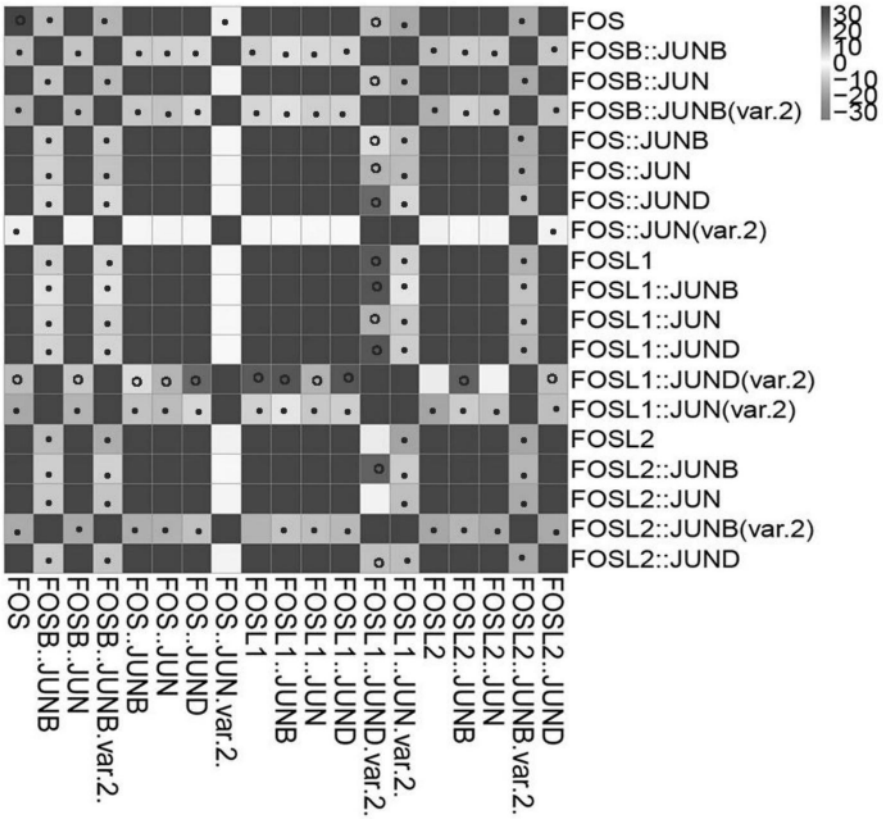

FOS
FOSB::JUNB
FOSB::JUN
FOSB::JUNB(var.2)
FOS::JUNB
FOS::JUN
FOS::JUND
FOS::JUN(var.2)
FOSL1
FOSL1::JUNB
FOSL1::JUN
FOSL1::JUND
FOSL1::JUND(var.2)
FOSL1::JUN(var.2)
FOSL2
FOSL2::JUNB
FOSL2::JUN
FOSL2::JUNB(var.2)
FOSL2::JUND
FOS
FOSB..JUNB
FOSB..JUN
FOSB..JUNB.var.2.
FOS..JUNB
FOS..JUN
FOS..JUND
FOS..JUN.var.2.
FOSL1
FOSL1..JUNB
FOSL1..JUN
FOSL1..JUND
FOSL1..JUND.var.2.
FOSL1..JUN.var.2.
FOSL2
FOSL2..JUNB
FOSL2..JUN
FOSL2..JUNB.var.2.
FOSL2..JUND
30
20
10
0
−10
−20
−30

图4C

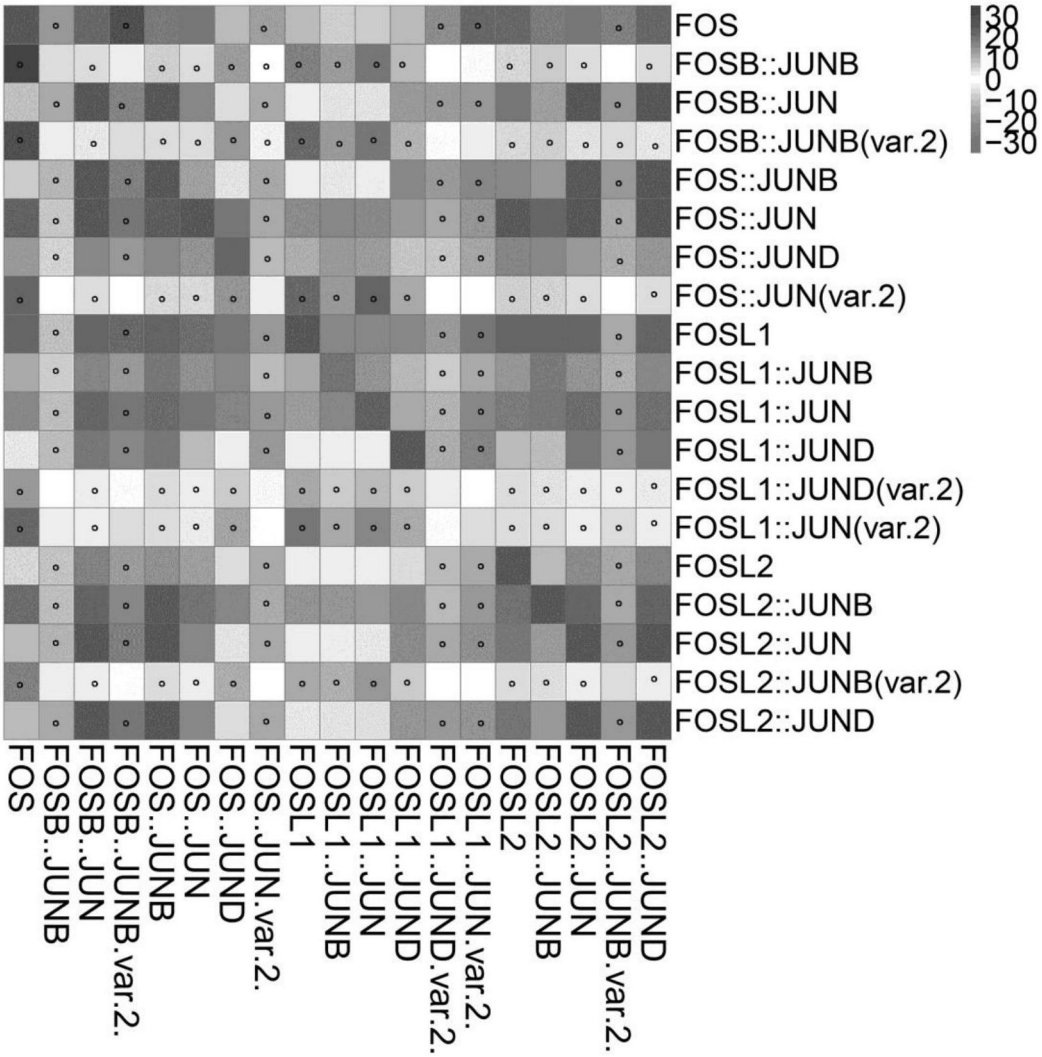

FOS
FOSB::JUNB
FOSB::JUN
FOSB::JUNB(var.2)
FOS::JUNB
FOS::JUN
FOS::JUND
FOS::JUN(var.2)
FOSL1
FOSL1::JUNB
FOSL1::JUN
FOSL1::JUND
FOSL1::JUND(var.2)
FOSL1::JUN(var.2)
FOSL2
FOSL2::JUNB
FOSL2::JUN
FOSL2::JUNB(var.2)
FOSL2::JUND
30
20
10
0
−10
−20
−30
FOS
FOSB..JUNB
FOSB..JUN
FOSB..JUNB.var.2.
FOS..JUNB
FOS..JUN
FOS..JUND
FOS..JUN.var.2.
FOSL1
FOSL1..JUNB
FOSL1..JUN
FOSL1..JUND
FOSL1..JUND.var.2.
FOSL1..JUN.var.2.
FOSL2
FOSL2..JUNB
FOSL2..JUN
FOSL2..JUNB.var.2.
FOSL2..JUND

图4D

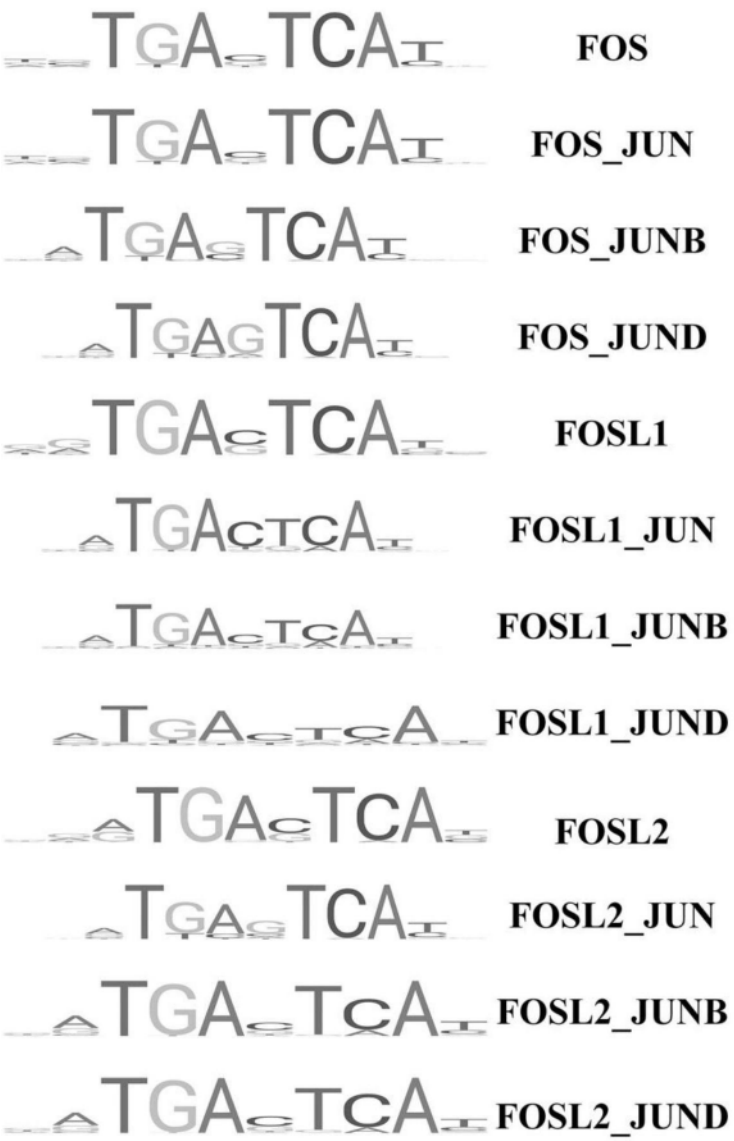

FOS
FOS_JUN
FOS_JUNB
FOS_JUND
FOSL1
FOSL1_JUN
FOSL1_JUNB
FOSL1_JUND
FOSL2
FOSL2_JUN
FOSL2_JUNB
FOSL2_JUND

图4E

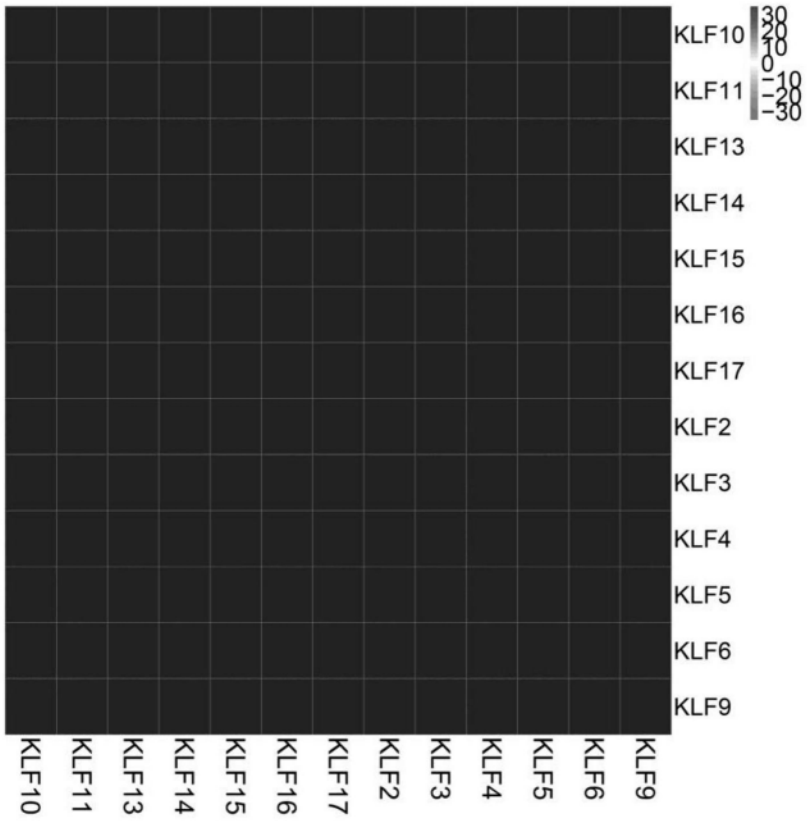

KLF10
KLF11
KLF13
KLF14
KLF15
KLF16
KLF17
KLF2
KLF3
KLF4
KLF5
KLF6
KLF9
30
20
10
0
−10
−20
−30

图5A

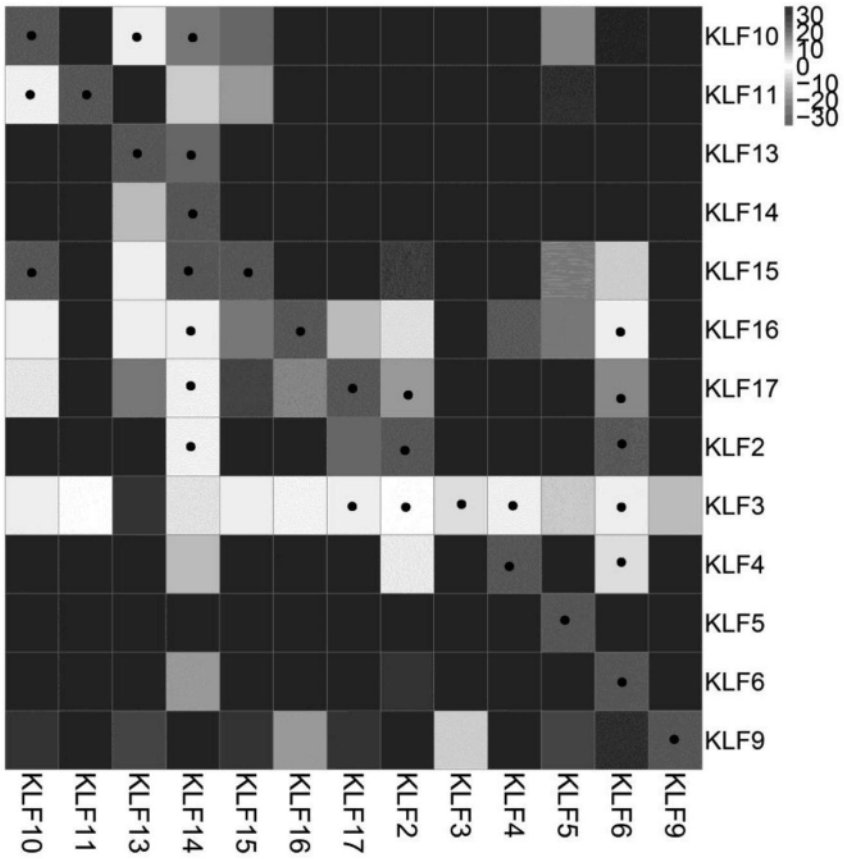

KLF10
KLF11
KLF13
KLF14
KLF15
KLF16
KLF17
KLF2
KLF3
KLF4
KLF5
KLF6
KLF9
30
20
10
0
−10
−20
−30

图5B

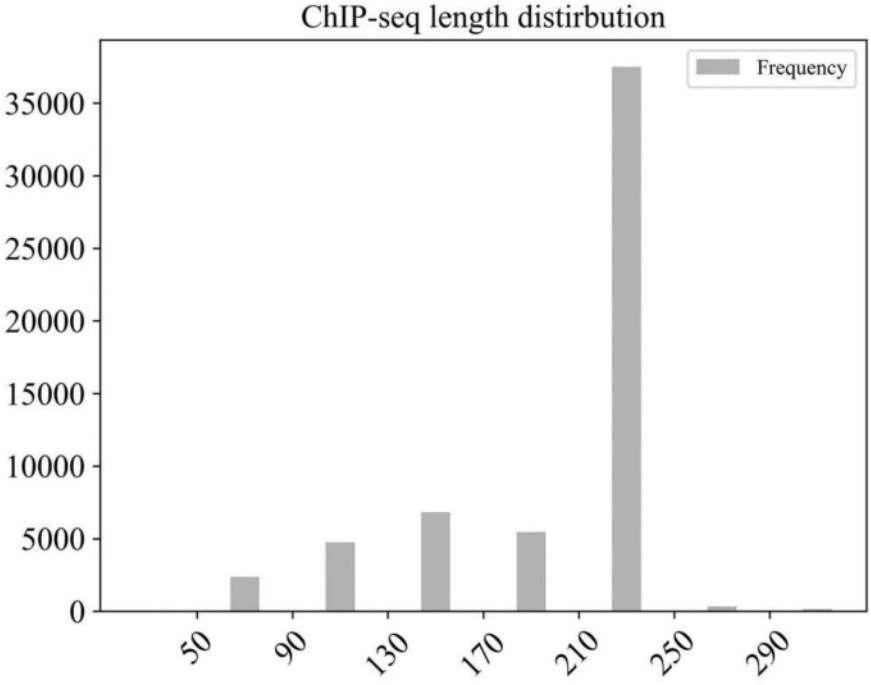

ChIP-seq length distirbution
Frequency
35000
30000
25000
20000
15000
10000
5000
0
50
90
130
170
210
250
290

图6A

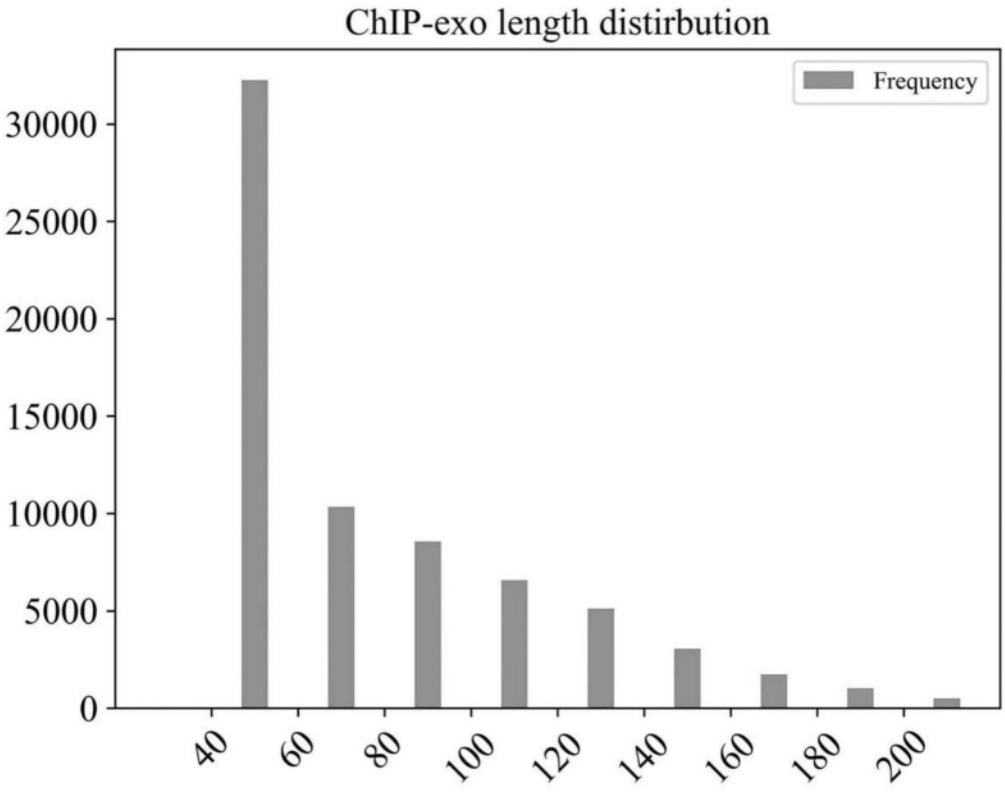

ChIP-exo length distirbution
Frequency
30000
25000
20000
15000
10000
5000
0
40
60
80
100
120
140
160
180
200

图6B

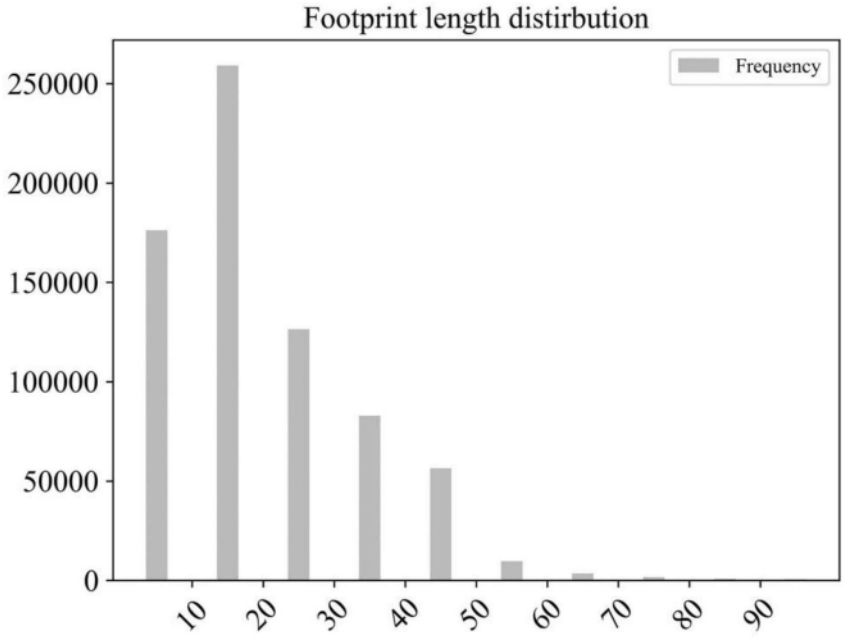

Footprint length distirbution
Frequency
250000
200000
150000
100000
50000
0
10
20
30
40
50
60
70
80
90

图6C

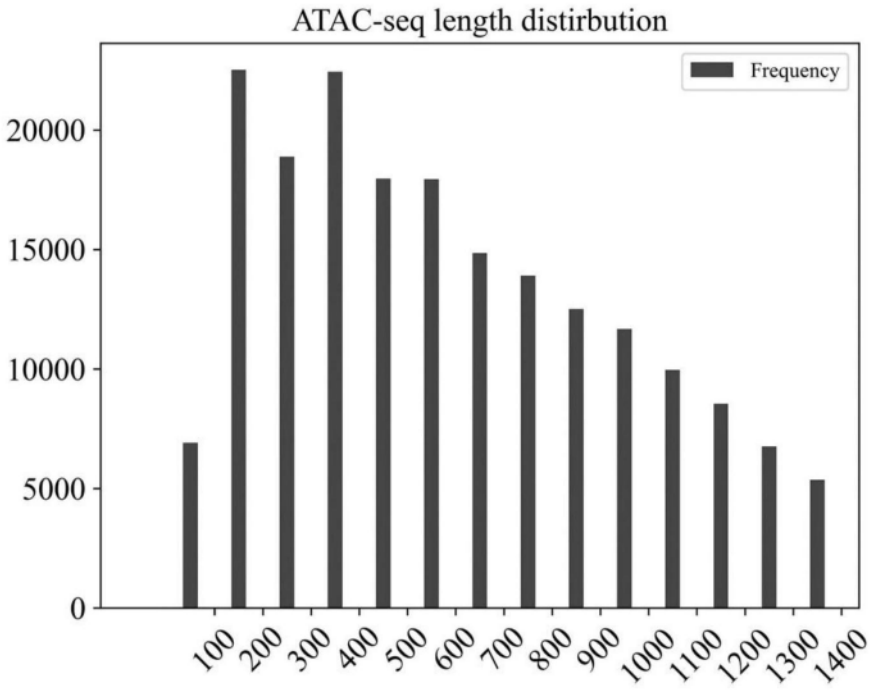

ATAC-seq length distirbution
Frequency
20000
15000
10000
5000
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400

图6D

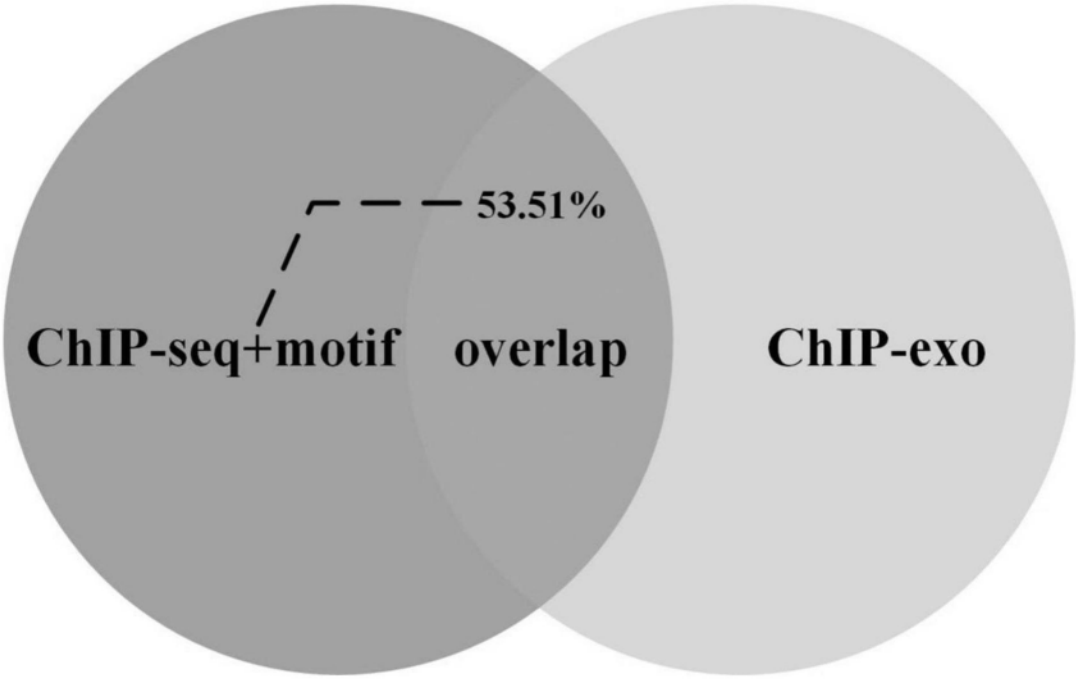

53.51%
ChIP-seq+motif
overlap
ChIP-exo

图7A

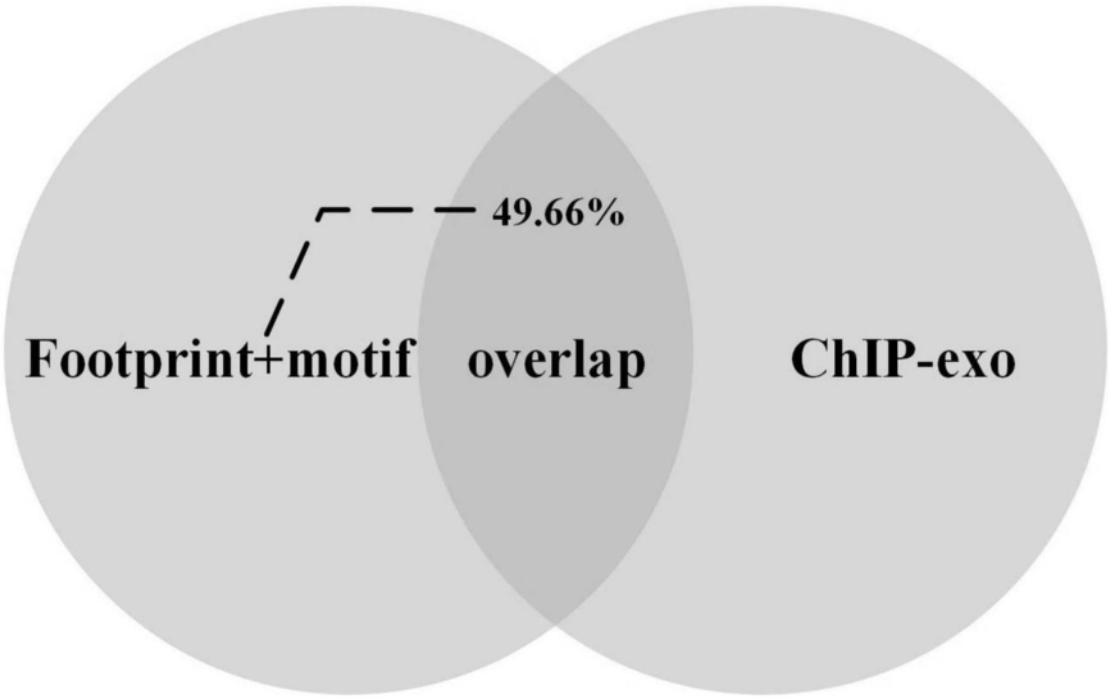


图7B