生物信息学基础知识和相关数据

Datalinkx

2025 年 4 月 13 日

1 转录过程

一个转录循环至少包括起始、延伸和终止 3 个主要阶段，其中转录的起始是最关键的调控机制，涉及各种转录调节因子同基因上游启动子序列之间的相互作用和结合、与聚合酶形成转录复合体、以及随后复合体的活化等一系列过程。在转录过程中，通常只复制一条 DNA 链。这被称为模板链，产生的 RNA 分子是单链信使 RNA（mRNA）。对应于 mRNA 的 DNA 链称为编码链。在真核生物（具有细胞核的生物）中，转录的初始产物称为 pre-mRNA。

真核生物有 3 种依赖 DNA 的 RNA 聚合酶 (RNA 聚合酶 1、2、3): (1)RNA 聚合酶 1 负责转录 rRNA; (2)RNA 聚合酶 2 负责转录结构基因 mRNA 和部分 snRNA; (3) RNA 聚合酶 3 负责转录 tRNA 和 5s rRNA。真核生物 RNA 聚合酶 2 必须与多与 20 个转录因子 TF2（transcription factor; 或称转录起始因子，transcription initiation factor，TIF；或称一般转录因子，general transcription factor，GTF）逐级组装成“完全的转录起始复合物”（complete transcriptional initiation complex，complete TIC）才能启动转录过程。

1. 转录的起始。起始阶段包括对双链 DNA 特定部位的识别、局部解链以及形成最初的一段 RNA。第一个核苷酸掺入的位置称为转录起点（transcription start site，TSS）。RNA 合成由 RNA 聚合酶催化。当 RNA 聚合酶与启动子区域结合时，转录就开始了。启动子包括转录成 RNA（起点）的第一个碱基对以及周围的碱基。启动子可以看作 DNA 上的一系列标志，指示出转录的起点、解开双螺旋的位点、RNAP 以及各种转录相关蛋白的结合位点等等。根据这些信息，转录才能顺利开始。

2. 转录的延伸。起始后 RNA 聚合酶（RNAP）构象改变，沿模板移动，持续合成 RNA，即进入延伸阶段。在转录启动时，Pol II 大亚基 Rpb1 的 C-末端结构域 (CTD) 第 5 位丝氨酸 (Ser5) 被磷酸化，PIC 部分解体，形成早期转录延伸复合体 (early transcript elongation complex, eTEC)，并进行早期转录延伸。但在新生 RNA 延伸至 20 30 nt 时，DSIF(DRB

sensitivity inducing factor) 和 NELF(negative elongation factor) 协同将 eTEC 阻滞于靠近启动子区的模板上，以便提供充足的时间为新生的 RNA 链 5'- 端加帽。此时，Ser5 磷酸化的 CTD 可结合 RNA 加帽相关的 3 个酶，依次对新生 RNA 进行加帽反应。当加帽完成后，停滞的转录延伸需要重新启动，于是转录延伸调控中最重要的正性转录延伸因子 b(positive transcription elongation factor b, P-TEFb) 被募集上来。它先分别磷酸化 DSIF 的 Spt5 亚基和 NELF 的 RD 亚基，执行解抑制的功能；该磷酸化的发生使 NELF 从 eTEC 上解离下来，并将 DSIF 逆转为促进转录延伸的因子。同时，P-TEFb 进一步磷酸化 CTD 第 2 位丝氨酸 (Ser2)，刺激 Pol II 的转录延伸活性，重新启动转录延伸，直至全长 mRNA 产物的形成。

3. 转录的终止。当聚合酶到达转录终点时，在终止因子的帮助下停止合成反应，酶和 RNA 链脱落，转录结束。

2 转录调控过程

2.1 真核生物染色质结构与基因活性

转录调控与染色质结构密切相关。活性增强子通常没有核小体结构，以利于 TF 的结合。而其附近的组蛋白经常具有表观遗传标记，例如 H3K4me1（组蛋白 H3 的 4 位赖氨酸单甲基化）和 H3K27 的乙酰化（H3K27ac）。而 H3K4 的三甲基化（H3K4me3）通常在基因启动子上富集。

- **组蛋白对基因转录活性的影响**

  1. 组蛋白与转录因子竞争基因的转录调控区。
  2. 未乙酰化的组蛋白可以抑制转录，乙酰化的组蛋白抑制转录的能力减弱；形成新的组蛋白共价键修饰（去甲基化），能抑制基因转录活性。
  3. 活跃转录的染色质区段，富合赖氨酸的组蛋白 (H1 组蛋白) 水平降低，组蛋白 2A /2B(H2A /H2B) 二聚体不稳定性增加、组蛋白发生乙酷化 (acetylation)、泛素化 (ubiquitination) 和组蛋白 3(H3) 统基化等现象，这些都是核小体不稳定或解体的因素或指征，转录活跃的区域也常缺乏核小体的结构。这些都表明核小体结构影响基因转录。

- **DNA 甲基化对基因转录活性的影响:** 真核 DNA 中的胞嘧啶约有 5% 被甲基化为 5 一甲基嘧啶 (5-methylcytosine，m5C)，而活跃转录的 DNA 片段中胞嘧啶甲基化程度常较低。这种甲基化最常发生在某些基因 5" 侧区的 CpG 序列中，实验表明这段序列甲基

化可使其后的基因不能转录，甲基化可能阻碍转录因子与 DNA 特定部位的结合从而影响转录。DNA 的甲基化对基因表达调控是重要的。

- **常染色质和异染色质:** 异染色质比常染色质压缩得更紧密，因此异染色质区的基因转录被抑制。
- **DNA 拓扑结构变化:** 天然双链 DNA 的构象大多是负性超螺旋。当基因活跃转录时，RNA 聚合酶转录方向前方 DNA 的构象是正性超螺旋，其后面的 DNA 为负性超螺旋。正性超螺旋会拆散核小体，有利于 RNA 聚合酶向前移动转录，而负性超螺旋则有利于核小体的再形成

2.2 转录激活因子对转录的影响

1. 顺式调控组件: 真核基因的顺式调控组件是基因周围能与特异转录因子结合而影响转录的 DNA 序列。其中主要是起正性调控作用的顺式作用组件，包括启动子 (promoter)、增强子 (enhancer)；近年又发现起负性调控作用的组件的沉寂子 (silencer)。

   (a) 真核基因启动子 (promoter) 是 RNA 聚合酶结合位点周围的一组转录控制组件。TATA box 是基本转录因子 TFIID 结合位点。

   (b) 增强子是远离转录起始点、决定基因的时间、空间特异性表达、增强启动子转录活性的 DNA 序列，其发挥作用的方式通常与方向、距离无关，可位于转录起始点的上游或下游。从功能上讲，没有增强子存在，启动子通常不能表现活性; 没有启动子时，增强子也无法发挥作用。

   (c) 沉默子 (silencer) 某些基因含有的一种负性调节元件，当其结合特异蛋白因子时，对基因转录起阻遏作用。某些基因有负性调节元件抑制子 (沉默子) 存在。有些 DNA 序列既可作为正性、又可作为负性调节元件发挥顺式调节作用，这取决于不同类型细胞中 DNA 结合因子的性质。

2. 反式作用因子: 以反式作用影响转录的因子可统称为转录因子 (transcription factors, TF)。RNA 聚合酶是一种反式作用于转录的蛋白因子

3 转录后调控

3.1 转录后加工

真核生物 mRNA 前体须经过 5'-加帽、3'-加尾以及拼接过程、内部碱基修饰才能成为成熟度的 mRNA，加帽位点与加尾位点、拼接点的选择就成了调控的手段。

- 5'-加帽：几乎所有的真核生物和病毒 mRNA 的 5' 端都具有帽子结构，其作用为保护 mRNA 免遭 5' 外切酶降解、为 mRNA 的核输出提供转运信号和提高翻译模板的稳定性和翻译效率。实验证实，对于通过滑动搜索起始的转录过程来说，mRNA 的翻译活性依赖于 5' 端的帽子结构。

- 3'-加尾：3'UTR 序列及结构调节 mRNA 稳定性和寿命

3.2 mRNA 稳定性

1. mRNA 的转录速度不变但稳定性增加，也可以增加蛋白质的表达量。

2. miRNA 与靶 mRNA 不完全互补，可导致 miRNA 在蛋白质翻译水平上抑制靶基因表达。miRNA 也有可能影响 mRNA 的稳定性。如果 miRNA 与靶位点完全互补（或者几乎完全互补），那么这些 miRNA 的结合往往引起靶 mRNA 的降解 (在植物中比较常见)。通过这种机制作用的 miRNA 的结合位点通常都在 mRNA 的编码区或开放阅读框中。每个 miRNA 可以有多个靶基因，而几个 miRNAs 也可以调节同一个基因。

   影响 mRNA 的稳定性有关的因素

   (a) 5'cap: 真核 mRNA 5' 末端帽结构的功用有二： 保护 5' 端免受磷酸化酶和核酸酶的作用, 从而使 mRNA 分子稳定； 提高在真核蛋白质合成体系中 mRNA 的翻译活性, 如果细胞内的脱帽酶 (decapping enzyme) 被 mR NA 中的序列元件激活, 则有可能导致 mRNA 的降解。因为 5'→3' 核酸外切酶, 或者某种作用位点曾被帽结构及与其偶联的结合蛋白所屏蔽的核酸内切酶, 此时便可乘虚而入, 对 mRNA 进行降解。

   (b) 5'UTR: 5' 非翻译区参与 mRNA 稳定性调控的证据不少, 对原癌基因的研究发现, 正常的 C-myc 基因的 mRNA 不稳定, 半衰期仅仅 0-15min。但突变的 C-myc 基因的 mRNA 被截短 (truncated) 了, 它们有正常的编码区, 3'UTR 及 poly (A), 却没有了通常的 5'UTR。但这截短了的 mRNA 的半衰期却比其正常的 mR NA 延长了约 3- 5 倍。

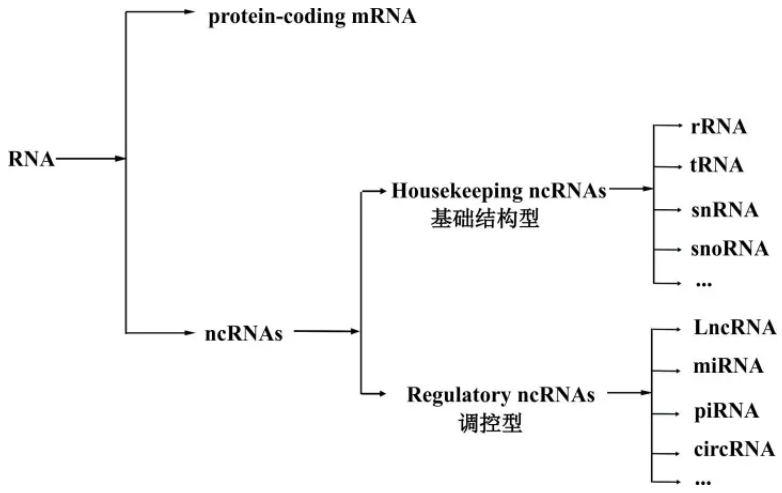

图 1: non-coding RNAs

(c) 编码区: 在编码区内有导致 mRNA 不稳定的序列 (称之为不稳定子, destabilizer) . 在哺乳动物细胞 C-fos mRNA 及 C-myc mRNA 的编码区中都发现了会促使 mRNA 降解的序列元件, 甚至已鉴定出一系列能识别这些不稳定性元件的结合蛋白。

(d) 3' 非翻译区 (3'U TR): 3' UTR 中最具普遍意义同时也是研究最为透彻的序列元件是稳定子 (stabilizer) - IR 序列形成的茎环结构, 和不稳定子- AU 富含元件 (ARE). 许多真核基因转录本的 3' 侧翼序列都可形成茎环结构。普遍认为该结构具有促进 mRNA 稳定的作用, 其主要依据是稳定的环结构既然能够阻碍反转录酶通过, 则同样可望抵御 3'→5' 核酸外切酶的降解活性, 从而在一定程度上加强了 mRNA 3' 末端的屏蔽作用。

(e) poly (A) 尾:poly (A) 尾的存在对提高 mRNA 稳定性具有促进作用, 不少研究已经证明了这一点。由于 poly (A) 尾缓冲了核酸外切酶对 RNA 3'→5' 方向的降解, 因此, po ly (A) 尾的缩短及其最终的脱离是 mRNA 降解的前奏。

3.3 non-coding RNA

图 1 展示了不同的 non-coding RNA 的分类。

不同的非编码 RNA 的功能简介

(a) **miRNA:** 在细胞核内, pri-miRNA 被 Drosha 加工成 pre-miRNA, 进而被 Exportin-

5 蛋白运输到细胞质。在细胞质中，pre-miRNA 被 RNase enzyme Dicer 加工，成为成熟的 miRNA

(b) **RNAi:** RNA 干扰（RNA interference, RNAi）是指在进化过程中高度保守的、由双链 RNA（double-stranded RNA，dsRNA）诱发的、同源 mRNA 高效特异性降解的现象。基因沉默分为转录水平的沉默（TGS）和转录后水平的沉默（PTGS）。

i. TGS 是指转基因在细胞核内 RNA 合成受到了阻止而导致基因沉默;

ii. PTGS 则是指转基因能够在细胞核里被稳定地转录, 但在细胞质里却无相应的 mRNA 存在这一现象。目前普遍认为 RNAi、共抑制、quelling 均属于 PTGS Dicer 核酸酶负责将 dsRNA 转化为 siRNA。siRNA 在细胞内 RNA 解旋酶的作用下解链成正义链和反义链,继之由反义 siRNA 再与体内一些酶 (包括内切酶、外切酶、解旋酶等) 结合形成 RNA 诱导的沉默复合物 (RNA-induced silencing complex，RISC)。siRNA 不仅能引导 RISC 切割同源单链 mRNA，而且可作为引物与靶 RNA 结合并在 RNA 聚合酶 (RNA-dependent RNA polymerase，RdRP) 作用下合成更多新的 dsRNA，新合成的 dsRNA 再由 Dicer 切割产生大量的次级 siRNA，经过若干次的合成-切割循环，RNAi 的作用不断放大，最终将靶 mRNA 完全降解。

(c) **LncRNA:**

(d) **piRNA:**

(e) **circRNA:**

4 翻译调控过程

1. mRNA 的选择性拼接
2. RNA 的编辑
3. RNA 的转运

4.1 翻译的起始调控

翻译起始过程中也需要多种蛋白因子，称为起始因子（initiation factor）。原核生物中以 IF 命名，真核称为 eIF（eukaryotic IF）。

1. 一些蛋白质可以结合 eIF4E，因而成为翻译阻遏物。研究最深入的翻译阻遏物是哺乳动物 eIF4E 结合蛋白（4EBP）。如果 4EBP 被高度磷酸化，就不能与 eIF4E 结合，从而失去翻译阻遏作用。

2. 在哺乳动物细胞中，mTORC1 复合物负责 4EBP 的磷酸化，在增殖和细胞生长中起关键作用。一旦 mTORC1 失活，就会发生 4EBP 的快速去磷酸化，恢复 eIF4E 结合能力，从而抑制了帽依赖翻译起始。在氧化应激、氨基酸限制、内质网应激等条件下，经典模式会受到一些通路的抑制，最常见的就是对 eIF2 和 eIF4E 的抑制。

4.2 mRNA 的选择性翻译

翻译起始因子与不同 mRNA 的亲和力不同，调节不同 mRNA 的翻译水平。

5 翻译后调控

蛋白质翻译后修饰 (Protein translational modifications，PTMs) 通过功能基团或蛋白质的共价添加、调节亚基的蛋白水解切割或整个蛋白质的降解来增加蛋白质组的功能多样性。这些修饰包括磷酸化、糖基化、泛素化、亚硝基化、甲基化、乙酰化、脂质化和蛋白水解，几乎影响正常细胞生物学和发病机制的所有方面。

6 Database

6.1 DNA

1. UCSC 基因浏览器，可以直接 download enhancer, promoter,UTR region, and so on.

2. Genotype: 全基因组关联分析（GWAS）是广泛用于寻找复杂遗传疾病关联基因的重要手段。其利用回归分析寻找染色体上的变异位点，并研究这些变异位点与疾病或其他性状的关联。

    - 人类
        - Europe: UK BioBank：Sample size=488,377；SNP=784,256; FinnGen: Sample size 343k; Estonian BioBank: Sample size 200k

  - Asia: Westlake BioBank for Chinese (WBBC): Sample size=14,726; China Kadoorie BioBank (CKB): Sample size 500k; Taiwan BioBank (TWB): Sample size 150; BioBank Japan (BBJ): Sample size 200k
  - Africa: Uganda Genome Resource: Sample size 6k
  - America: Michigan Genomics Initiative: Sample size 55k; Penn Medicine Biobank: Sample size 40k; UCLA Precision Health Biobank: Sample size 27k
- GWAS Summary statistics:
  - GWAS ATLAS 该数据库是公开的 GWAS 摘要统计数据的综合数据库。在该网站中，用户不仅可以访问原始汇总统计数据，还可以从预先执行的分析中获得各种结果，例如风险位点信息、LD 评分回归、MAGMA 和多 GWAS 比较。该数据库目前包含 4756 个 gwas 数据，来自 473 个不同的研究中的 3302 个特征，覆盖了 28 个区域。该数据库只提供了基于浏览器的下载，没有提供函数式 API 接口。
  - GWAS Catalog 该数据库于 2008 年由 National Human Genome Research Institute (NHGRI) 建立，旨在应对快速增长的全基因组关联分析（GWAS）。该数据库中主要包含了已发表 GWAS 的 Summary statistics. 截止 2023 年 7 月 25 日，该网站中共收录了 58,302 条数据，每条数据都有一个唯一的 GCST accession ID 与之对应。该数据库提供了浏览器端基于 ftp 的下载，同时提供了基于 HTTP 对话的 API 接口。
  - OpenGWAS 该数据库由布里斯托大学 MRC 综合流行病学部门 (IEU) 开发，是手动整理的 GWAS 摘要数据集，可作为开源文件下载，或通过查询完整数据的数据库来获取。该数据库提供的 API 可用来获取 GWAS summary 数据，并进行一些相关的分析，可通过相应的 R 语言或 python 的包实现数据访问和分析。
- 动物: mouse: Sample size=1,150；SNP=92,734; sheep: Sample size=91；SNP=45,342; pig: Sample size=2,150； SNP=36,740; chicken : Sample size=163； SNP=47,728.
- 植物：cucumber: 样本数 =836； SNP 个数 =23,552; cotton: 样本数 =258； SNP 个数 =1,871,401

3. 基因组 (除人以外的其他物种，比如昆虫、花卉、动物等)

4. DNA elements

- ENCODE 包含：人、老鼠、蠕虫、苍蝇这四个物种

- Enhancer (增强子): 增强子作为基因组上的顺式作用元件, 其主要作用是增强目标基因的表达。科学家提出了超级增强子 super-enhancer (SE) 的概念，将基因组上富集了增强子的区域定义为超级增强子
  - 超级增强子数据库
  - TiED: 人类增强子数据库，对 10 种不同组织中的增强子表进行定量和分析，鉴定了增强子的组织特异性
  - 人和小鼠中的超级增强子数据库. 对于 human 而言，基于 hg19 版本进行分析，共收录了来自 102 种不同细胞/组织共 69205 个超级增强子; 对于 mouse 而言, 基于 mm9 版本进行分析，共收录了来自 25 种不同细胞/组织共 13029 个超级增强子信息
  - 如果知道 enhancer 的位置, 想知道其对应的基因, 但目前没有完备的数据库, 因为每个细胞的情况都不一样, 一般是根据距离直接注释到基因上, 再分出来 promoter 和 enhancer, 但是因为同一个 enhancer 大约会调控 1.6 个基因, 同一个基因大约会被 5.53 个元件调控
- 启动子 (promoter)
  - Eukaryotic Promoter Database: 比较全的基因启动子数据，并且经过试验验证；同时数据库给出 UCSC 中 gene 起始位点的数据下载链接
  - MPromDb: 收集小鼠和人类的启动子，注册后可下载数据
  - TFBIND: 结合 TRANSFAC 数据库预测 gene 的启动子，使用的 TRANSFAC 版本为 TRANSFACR.3.4
  - AtProbe: 拟南芥启动子结合元件数据库，给出的元件有 118 个、基因 76 个
- 转录因子 (TF): 转录因子，也称为序列特异性 DNA 结合因子，是一群能与基因 5'端上游特定序列专一性结合，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的蛋白质分子。真核生物在转录时往往需要多种蛋白质因子的协助。一种蛋白质是不是转录机构的一部分往往是通过体外系统看它是否是转录起始所必须的。
  - iRegulon包含转录因子及其直接转录 DNA 序列组成，在转录序列的顺式作用元件处包含与 TF 结合位点。
  - ORegAnno(The Open Regulatory Annotation Database) is a dynamic collection of literature-curated regulatory regions, transcription factor binding sites and regulatory mutations. (polymorphisms and haplotypes)
  - checkpoint是人类、小鼠和大鼠转录因子数据库。手动检索 TFcheckpoint 中的转录因子可以获得其在 RNA 聚合酶 II 调节和特异性 DNA 结合活性实验中的数据。

- TSS: DBTSS 7 个物种（人类、小鼠、大鼠、疟原虫、红藻、穴居人、猴）的转录起始位点预测，这些 TSS 位点是经过 TSS-seq 实验验证的。可以查看 TSS 区域的 SNV 信息

6.2 DNA regulation

1. DNA-DNA interaction: Liebermann-Aiden 等在 2009 年开发出基于高通量测序方法在全基因组范围内研究染色质空间构象的新技术 Hi-C。Hi-C 可以来观察三维基因组。Hi-C 文库一次可以获取全基因组范围内互作的染色质片段，可以从全基因组的高度来研究染色质的空间结构特征。

   - 人类基因组 Hi-C 数据
   - 3CDB是一个染色质空间互作的数据库，根据特定的关键词从 pubmed 数据库中进行文献检索，查找基于 3C 技术研究染色质互作的文献，并从中提取染色质互作信息。收录了来自 17 个物种，共 3319 个染色质片段互作信息

2. DNA-protein(TF)

   - Chip-seq: 染色质免疫沉淀、DNase-I 超敏反应和转座酶可及性分析结合高通量测序，使染色质动力学、转录因子结合和基因调控的全基因组研究成为可能。全基因组范围内的转录因子（TF）和染色质调节剂结合位点，组蛋白修饰和染色质可及性的鉴定对于控制生物过程（如分化，致癌作用和细胞对环境干扰的反应）的常规转录调控至关重要。
     - chipBase 收集来自 GEO,ENCODE 数据库中的 chip-seq 数据，通过对这些原始数据进行分析，致力于构建各种转录因子与非编码 RNA, 蛋白编码基因之间的调控网络
     - Cistrome DB 收集了 2016 年 1 月 1 日之前发布的 ChIP-seq 和染色质可及性数据（DNase-seq 和 ATAC-seq），包括 13366 人和 9953 小鼠样品。Cistrome DB 更新版于 2018 年发布，包含大约 47000 个人类和老鼠样本，与第一版相比，新收集了大约 24000 个数据集。
   - ATAC-seq: TCGA 数据库; ATAC-Seq 数据的下载
   - DNase-Seq

3. Regulatory network inference. ARACNe, WGCNA, statistical model

6.3 RNAs

1. mRNA

   - Bulk RNA-seq
     (a) TCGA: TCGA 数据库纳入的正常组织测序结果是非常少的，缺少病人的正常组织的转录组测序结果
     (b) GTEx 数据库包含了来自健康人的多种不同组织器官的基因表达数据 (组织特异性基因表达和调控数据库); 还有 snp 信息 (eQTL), 寻找基因之间和个体之间基因表达的差异; 要把 GTEx 数据库的转录组表达矩阵和 TCGA 的进行比较，还需要一定程度的去除批次效应。
   - single-cell RNA-seq
     (a) Human
     – easybioai 包括人体各种组织的单细胞数据各种疾病的各种细胞类型准确的基因表达谱资源，浏览感兴趣的基因的表达、搜索细胞类型标记、寻找多种疾病的生物标志物、比较处于疾病和非疾病状态的各种类型细胞的表达谱。
     – Human Cell Atlas是人类细胞图谱 (HCA) 计划，是重大国际科学合作计划，旨在全面解码人体所有细胞的类型、数目、位置、相互关联与分子组成等，构建细胞基因表达等高维数学特性的精细图谱，建成人体发育、生理、病例的完善和精细的参照系，最终建立全息生命信息网络。现已收录 34 个组织、295 位供体、450,0000 个单细胞的测序数据，并在持续更新中。HCA 的界面操作简单，用户可以直接在 “组织” 界面，点击相关组织（如: blood），进入到数据存放界面。该界面以列表形式呈现，主要是数据的一些基本信息（项目名，物种类型，测序平台等）。点击 “Export Selected Data” 按钮即可下载相应的数据。
     – JingleBells 将单细胞数据划分为免疫和非免疫类，共收录 120 篇与免疫相关以及 182 篇非免疫的单细胞文献，且数据库仍在持续更新中。用户在 JingleBells 上可以直接下载到单细胞数据的 BAM 文件。
     – CancerSEA 第一个专门解析癌症单细胞状态图谱的数据库。该数据库涉及到 25 种癌症类型的 4,1900 个癌症细胞的 14 个功能状态。用户可以点击主页上的 “download” 按钮进入下载界面，但是在该界面中用户只能下载到单细胞数据集的功能状态配置文件，而原始测序数据用户必须通过 GEO accession 进入 NCBI 中下载。此外，该数据库使用的数据集均为 2018 年 7 月份之前发表的数据，并未收录最新的高通量单细胞数据。

- HCL - Human Cell Landscape (zju.edu.cn) 70W 个细胞，100 余种细胞大类，800 余钟细胞亚类。
- scRNASeqDB (uth.edu) 包含了目前几乎所有可用的人类单细胞转录组数据集（n=38），覆盖 200 个人类细胞系或细胞类型和 13440 个样本。
- 欧洲 EMBL-EBI，收录了各种疾病类型的单细胞数据，包括 18 个物种、229 个 study、597 万个细胞。可以按 gene 和 experiment 检索实验设计、分析参数、下载 marker 基因和表达数据矩阵等。
- CancerSEA 包含 25 种癌症的 41900 个肿瘤细胞，14 种癌症相关功能状态，允许用户查询基因 (包括 PCGs 和 lncrna) 与 14 种功能状态之间的关系
- CancerTracer主要提供两类数据，分别是肿瘤内或转移内异质性: 原发肿瘤或单个转移病灶内存在多个亚克隆和转移后异质性: 同一患者不同转移灶存在不同亚克隆

(b) mouse

- Mouse Cell Atlas 1,130,000 single cells from >10 mouse tissues (2-4 replicates per tissue in general) at ten life stages (E10.5, E12.5, E14.5, Neonatal, 10d, 3w, 6 8w, 12m, 18m, 24m) from early embryonic stage to the mature adult stage.
- tabula-muris 小鼠整个生命周期内 20 个器官的 529823 个单细胞的 RNA 测序数据，使用两种方法分离出 100,000 多种细胞类型进行功能注释

(c) 多物种数据库

- 植物、动物、微生物: 提供了很多单细胞数据的注释和下载地址，scRNA-seq 数据检索/注释数据库下载文件为.h5 格式包括各种等
- 人和老鼠 小鼠组织 184 种、样本 1063 个、近 446W 个细胞；人类组织 74 种、样本 305 个、细胞 112w+。同时包含了 6000 多个 marker 基因，可用于细胞分群注释的 marker 数据库
- Single Cell Portal (broadinstitute.org) 收录 419 个研究、超过 1934 万细胞的单细胞数据库
- 人类以及动植物: 7015 个样本和 1775570 个细胞
- 人类和小鼠: 单细胞转录组学全面的元数据和转录组数据以及细胞图像信息
- cedr: 超过 582 个人类、小鼠和细胞系的单细胞数据结果，包括约 140 个表型和 1250 个组织/细胞类型的约 188,157 个人类相关、42660 个小鼠相关和 10299 个细胞系相关的细胞药物反应信息。

- 免疫相关的单细胞数据库
- 人和小鼠的脑血管、肺血管的单细胞数据
- 肾脏组织的单细胞数据
- 14 例非小细胞癌 初治患者外周血、癌组织和癌旁组织的 12,346 个 T 细胞的单细胞测序

(d) 相关分析

- sc-lncRNAs 比较分析单细胞 RNA-Seq 数据中 lncRNAs 表达、分类和功能的数据库
- 调控网络 分析转录因子 (TFs) 和下游靶基因 (称为调控子) 在各种组织/条件下形成的调控网络。

2. miRNA

- mirdb 用于 miRNA 靶标预测和功能注释的在线数据库。通过分析高通量测序实验中的数千个 miRNA-靶标相互作用而得到的数据人、小鼠、大鼠、狗和鸡五种动物（有一部分数据是预测得到的）
- mir2diesease 提供各种人类疾病中 miRNA 失调的数据, 包括 miRNA ID，疾病名称，miRNA-疾病关系的简要描述，疾病状态下的 miRNA 表达模式，miRNA 表达的检测方法，实验验证的 miRNA 靶基因以及文献参考.Number of miRNAs: 349，Number of diseases: 163，Number of entries: 3273
- mirbase 已发表的 miRNA 序列和注释的可检索数据库

3. tRNA: 包含来自 577 个物种超过 12000 个 tRNA 基因和来自 104 个物种的 623 个 tRNA 序列

4. lncRNA: lncRNAdb是人工整理的具备生物学功能证据的 lncRNA 数据库。因此 lncRNAdb 的资料可信度很高

5. NONCODE 数据库是一个综合的非编码 RNA 数据库，该数据库中包含了除 tRNA 和 rRNA 之外的其他类型的非编码 RNA 信息，其中绝大部分是 lncRNA. 目前最新版本为 v5, 共包含了 17 个物种的非编码 RNA

6. ceRNA 调控网络: TargetScan, miRDB, miRTarBase, miRWalk 预测 miRNA-mRNA 靶向关系; ceRNA 是 miRNA, lncRNA, circRNA 等的统称。

6.4 Proteins

1. MS data is the key kind of protein sequence.
2. Protein-Protein interaction database: string
3. gene-protein: BioGRID
4. Post-translational modification database
5. PubChem 有机小分子生物活性数据库
6. Protein acitivity inference. metaVIPER.
7. Receptor-ligand interaction.
8. Transcription factors (TF): TF-target
9. UniProt 由 EBI(欧洲生物信息研究所)、PIR(蛋白信息资源) 和 SIB(瑞士生物信息研究所) 合作建立而成，提供详细的蛋白质序列、功能信息，如蛋白质功能描述、结构域结构、转录后修饰、修饰位点、变异度、二级结构、三级结构等，同时提供其他数据库，包括序列数据库、三维结构数据库、2-D 凝聚电泳数据库、蛋白质家族数据库的相应链接。其中的 GO 数据文件可见附录 uniprotdata_go.zip, 其中有五个 csv 文件，是所有蛋白质按序列长短分为了五个文件，下标越小序列越短。
10. BRENDA 酶数据库，起源于德国不伦瑞克在 1987 年建立的国家生物技术研究中心 (GBF)，目前由德国科隆大学生物化学研究所负责运营。BRENDA 可以提供酶的分类、命名法、生化反应、专一性、结构、细胞定位、提取方法、文献、应用与改造及相关疾病的数据。
11. InterPro 蛋白质综合数据库，从大量的数据库中整合而成的包括蛋白质结构域、蛋白质家族、功能位点等信息的数据库。interproscan 蛋白质结构域预测。
12. ConsensusPathDB 是一个分子功能相互作用数据库，集成了有关人类蛋白质相互作用、遗传相互作用信号、代谢、基因调控和药物靶点相互作用的信息。
13. Reactome
14. PID
15. KEGG

16. PDB 生物大分子结构数据库，提供蛋白质、核酸等生物大分子的三维结构数据、序列详细信息、生化性质等。

17. UniHI 人体蛋白-蛋白相互作用数据库，可根据蛋白质名称、代谢路径等进行查询。

18. STITCH 蛋白质-化合物作用网数据库

19. DOMINE 结构域互作数据库。

20. 3DID 搜集 3D 结构已知的蛋白质的互作信息，可通过结构域名称、基序名称、蛋白质序列、GO 编码、PDB ID、Pfam 编码进行检索。

21. STRING 蛋白质互作网络数据库

22. CELLO 蛋白质亚细胞结构定位

23. MFUZZ 蛋白质聚类分析

24. HALLMARK 蛋白质功能标注，one-hot 编码文件（由.gmt 文件整理而来）见附件 hallmark.csv，其中蛋白质使用 NCBI ID 编号。

25. Human Protein Atlas: 人类蛋白质图谱是一个综合性数据库，提供了人类蛋白质在组织和细胞水平上的表达信息，包括免疫组化、蛋白质亚细胞定位等。

6.5 Pathway

生物通路是细胞内分子之间的一系列相互作用，会导致细胞内产生某种产物或某种改变。通路可以触发新的分子如脂肪或蛋白质的组装。通路也可以开启或关闭基因，或者刺激细胞移动。最常见的生物通路涉及到代谢、基因表达调控和信号传导。通路信息可以从大量数据库获得，包括从专业校对过的高质量数据库和通过对文章摘要进行自然语言处理和文本挖掘产生的大量假定通路的数据库。

1. pathway 的数据库

   (a) 人类

      i. KEGG pathway

         - 新陈代谢: 包括碳水化合物代谢、能量代谢、脂质代谢、核苷酸代谢、氨基酸代谢等相关通路；
         - 遗传信息加工: 包括转录、翻译、折叠、分选、降解、复制和修复等相关通路；

- 环境信息加工: 包括膜转运、信号转导、信号分子相互作用等相关通路;
- 细胞过程方面: 包括运输和代谢、细胞生长和死亡、细胞群落、细胞运动等相关通路;
- 生物体系统方面: 包括免疫系统、内分泌系统、循环系统、消化系统、神经系统等相关通路;
- 人类疾病方面: 包括肿瘤、免疫性疾病、神经变性疾病、心血管疾病、内分泌代谢性疾病、药物抗性等相关通路;
- 药物开发方面: 包括抗感染药、抗肿瘤药、神经系统药物等相关通路; 另外还包含靶向药物的相关通路, 比如 G 蛋白偶联受体通路、核受体通路、离子通道通路、转运蛋白通路、酶通路等。

ii. BioCyc:15,037 Pathways. 第一层数据库包含 EcoCyc、MetaCyc 和 BOCD; 第二层和第三层数据库则包含了计算预测代谢通路, 以及哪些基因编码代谢通路中缺少酶的预测和预测的操纵子。单个生物体的基因组和代谢途径; 代谢通路上游的基因调控信息; 代谢通路与基因编码的酶及其调节因子之间的联系. BioCyc 提供的分析工具包括: 基因组浏览器、个体代谢通路和完整代谢图的显示等。

iii. Reactome:26857 pathways 直观的可视化生物信息学工具; 以人类相关数据为主, 同时包含 22 种其他物种的数据, 比如小鼠和大鼠; 包括代谢通路、信号转导、基因转录调控、细胞凋亡与疾病等;

iv. PathBank 110,315Pathways PathBank 是一个交互式的可视化数据库, 包含在人类、小鼠、大肠杆菌、酵母和拟南芥等模型生物中发现的 10 万多条机器可读路径。大多数这些途径在任何其他途径数据库中都找不到。PathBank 专门设计用于支持代谢组学、转录组学、蛋白质组学和系统生物学中的通路阐明和通路发现。所有 PathBank 路径都包括有关细胞器、亚细胞区室、蛋白质复合物辅因子、蛋白质复合体位置、代谢物位置、化学结构和蛋白质复合物四级结构的信息。

v. BIOCARTA 用来进行分子互作关系、富集分析、通路为基础的研究

vi. XTalkDB 研究信号通路间相互作用的数据库

vii. WikiPathways: 1,946Pathways 由科学界维护并为科学界服务的生物通路数据库

viii. Pathway Interaction Database: 745 Pathways。PID 是人类细胞信号通路的传统生物医学数据库。包含有关细胞中发生的分子相互作用和反应的精选信息, 特别关注可能与癌症研究和治疗相关的过程。

ix. PathCommons: 综合了大部分的通路数据库: 5772 Pathways – 2424055 Interactions – 22 Databases

(b) plant

- PlantCyc: 64,581Pathways 植物代谢通路数据库提供了对超过 500 种植物中存在的共享和独特代谢途径的手动整理或审查信息的访问。
- Plant Reactome:18,910Pathways 一个可自由访问的植物代谢和调节途径数据库。、目标是为植物研究人员提供可视化、解释和分析途径知识的工具，以支持基础研究、基因组分析、建模、系统生物学和教育。

(c) others: INOH：511Pathways. INOH 数据库是一个高度结构化的、人工策划的信号转导途径数据库，包括哺乳动物、非洲爪蟾、黑腹果蝇、秀丽隐杆线虫和典型。

(d) 微生物

与 pathway 相关的生物学分析

- 富集分析
  - GO(Gene Ontology) 由基因本体论联合会建立, 该数据库将全世界所有与基因有关的研究结果进行分类汇总。对不同数据库中关于基因和基因产物的生物学术语进行标准化, 对基因和蛋白功能进行统一的限定和描述。利用 GO 数据库, 可以在以下三个方面对基因和基因产物进行分类注释。BP：Biological Process, 生物过程 MF：Molecular Function, 分子功能 CC：Cellular Component, 细胞组分在这三个大分支下面又分很多小层级（level），level 级别数字越大，功能描述越细致。最顶层的三大分支视为 level1，之后的分级依次为 level2，level3 和 level4。通过 GO 注释，可以大致了解某个物种的全部基因产物的功能分类情况。GO 定义的术语具有有向无环性（directed acyclic graphs ,DAGs) 的特点，而并非是传统的等级制定义方式（随着代数增加，下一级比上一级更为具体）。
  - 京都基因与基因组百科全书(Kyoto encyclopedia of genes and genomes, KEGG) 是系统分析基因功能、基因组信息的数据库，整合了基因组学、生物化学及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体进行研究。目前 KEGG 共包含了 19 个子数据库，富集分析常用在 KEGG Pathway 通路中。
  - MSigDB 基因集富集: MSigDB 数据库定义了已知的基因集合，包括 H 和 C1-C7 八个系列 (Collection)。H: hallmark gene sets（效应) 特征基因集合，共 50 组；C1: positional gene sets 位置基因集合，根据染色体位置，共 326 个；C2:

curated gene sets：（专家）共识基因集合，基于通路、文献等，包括 KEGG；C3: motif gene sets：模式基因集合，主要包括 microRNA 和转录因子靶基因两部分；C4: computational gene sets：计算基因集合，通过挖掘癌症相关芯片数据定义的基因集合；C5: GO gene sets：Gene Ontology 基因本体论；C6: oncogenic signatures：癌症特征基因集合，大部分来源于 NCBI GEO 未发表芯片数据；C7: immunologic signatures: 免疫相关基因集合。可以从中获取大量的已知基因集合从而进行富集分析。

  - 单基因富集: 基于单个基因来抓取与其相关的基因，然后用这些相关的基因来进行功能富集，有两种方法：差异法和相关法。差异法：根据给定的一个基因的表达值对样本进行分组，然后计算组间的差异表达基因，进而利用差异基因进行富集分析; 相关法：计算给定的一个基因的表达值与其他基因之间的相关性，将具有显著相关的基因作为一个集合进行富集分析。

- pathway 活性分析

6.6 代谢组数据库

代谢组是指生物体内源性代谢物质的动态整体。而传统的代谢概念既包括生物合成，也包括生物分解，因此理论上代谢物应包括核酸、蛋白质、脂类生物大分子以及其他小分子代谢物质。通常我们需要了解代谢物的名称、CAS 号、物理化学性质、相关功能，以及与其他代谢物或者基因、酶间可能的相互作用等。

1. 数据库

   - Pubchem 由美国国家健康研究院（US National Institutes of Health，NIH）创建，主要收录有机小分子物质，包括化合物的化学结构标识符、化学和物理性质、生物活性、专利、毒性数据等许多信息。输入 ATP 进行检索，点击后可查询 ATP 的结构、物理化学性质、功能、用途等各项信息。
   - ChemSpider 是一个免费的化学结构数据库，提供对来自数百个数据源的超过 1 亿个代谢物结构的快速文本和结构搜索访问。
   - CHEBI 是一个收录生物医学相关化学条目的数据库，属于 EBI（欧洲生物信息研究所 (European Bioinformatics Institute)）旗下代谢物数据库。
   - KEGG 是代谢组常用数据库。在 KEGG compound 中检索 ATP，之后点击 id 号 C00002 可查看 ATP 各种信息以及可发生的反映和信号通路等信息，如需进一步探究与其他代谢物、基因的调控关系，可点击 Reaction、Pathway 查看。

- HMDB 中输入 ATP 检索，可得到包括 ATP 的命名、结构、解释、功能、应用等相关信息。（对于小鼠来说不能和 KEGG 互通且较少）
- Lipidmaps 是脂类物质最大最权威的综合性数据库，对于脂类物质的信息查询很有帮助。

6.7 表型组

- 人类表型组数据

1. The Human Phenotype Ontology
2. OMIM 是人类基因和遗传表型的全面、权威的数据库，也称：人类孟德尔遗传在线数据库。
3. Clinvar 是 NCBI 临床突变数据库，整合遗传变异、临床表型、支持证据以及功能注解与分析四方面的信息，采用星标系统来评价特定突变在疾病中的功能注释等级，记载文献中变异与疾病/表型之间的关系，且有文献溯源。
4. SwissVar 数据库 Swiss 汇总了与特定变体有关的所有信息，并包含：根据文献对每种特定变体的基因型-表型关系进行人工注释；预先计算的信息（例如保护评分和可用的结构特征列表），以帮助评估变体的效果。包括疾病信息、蛋白信息、结构或功能特征信息。
5. MalaCards 是人类疾病综合性数据库，参考 GeneCards 数据库的架构，整合了专业和一般疾病，包括罕见疾病、遗传疾病、复杂疾病等。
6. EGA包含多种测序以及分型数据，如基因组关联分析、分子诊断以及各种目的的测序数据。约 60% 表型都与肿瘤相关。
7. ICGC癌症数据库，包括 50 种不同癌症类型/亚型的肿瘤中的基因组异常（体细胞突变，基因异常表达，表观遗传修饰）数据。

- 细胞表型数据

1. The Cell Image Library (CIL): 是一个关于细胞成像数据的在线资源，包含各种细胞类型的显微镜图像和视频，用于研究细胞形态和功能。
2. CellMiner: 提供了大量肿瘤细胞系的分子特征、基因表达数据和药物敏感性信息，有助于药物治疗和细胞生物学研究。
3. Organelle Genome Resources: 这些资源提供了不同细胞器 (如线粒体、叶绿体) 基因组的数据，有助于研究细胞器功能和进化。

4. Allen Cell Explorer: 该资源提供了大量细胞图像和数据，用于研究细胞的形态、结构和功能，包括三维细胞成像。
5. EMDataBank: 该数据库收集和存储电子显微镜 (EM) 图像和结构数据，包括细胞、蛋白质复合物等的高分辨率结构信息。
6. CellSys: CellSys 提供了关于细胞生物学、细胞信号传导、细胞分裂等方面的模型和数据，用于研究细胞的动态过程。
7. CellPhoneDB: 该数据库提供了细胞-细胞相互作用的信息，有助于理解细胞间的通讯和信号传递。
8. The Cell Ontology (CL): 细胞本体是描述细胞类型和细胞组分的资源，有助于标准化细胞表型组学数据的表示和共享。

- 细胞注释相关

1. CellMarker: 是一个用于细胞类型鉴定的数据库，提供了不同细胞类型的标记基因信息。
2. B 细胞数据库
3. 正常人及 AML 患者、正常鼠的各种血细胞类
4. Human Cell Atlas: 人类细胞图谱计划旨在创建一个关于人类体内所有细胞类型的细胞图谱，提供关于细胞组成和功能的全面信息。
5. Allen Brain Atlas: 该资源提供了关于大脑组织中基因表达的空间信息，有助于理解大脑的细胞类型和功能。

- 在线分析工具

1. CellProfiler: CellProfiler 是一个开源的图像分析工具，用于高通量细胞成像数据的分析和处理。
2. CellFishing.jl: 这是一个用于处理高维细胞数据的 Julia 软件包，有助于细胞表型组数据分析和解释。
3. BioGPS: BioGPS 是一个基因注释和表达数据的在线平台，可以帮助用户查找不同组织和细胞类型中基因的表达信息。
4. Cytoscape: Cytoscape 是一个用于分析和可视化细胞网络和通路的工具，有助于理解细胞内部的相互作用关系。
5. CellXgene: 该工具用于单细胞 RNA 测序数据的交互式分析和可视化，有助于探索单细胞基因表达模式。

6. ExPASy - Proteomics Server: ExPASy 提供了与蛋白质组学相关的工具和资源，有助于研究蛋白质组的表型和功能。

- 药物相关数据库

1. Therapeutic Target Database, TTD: 包含已知和研究中的治疗性蛋白质和核酸靶点、疾病、通路信息以及相应的药物信息。
2. The DrugBank Database, DrugBank: 该数据库整合了详细的药物 (即化学、药理和药剂) 数据与全面的药物靶点 (即序列、结构和途径) 信息。
3. Binding Database, BindingDB: 专注于收集药物靶点蛋白质和类药小分子之间相互作用亲和力的数据。这一资源有助于研究者通过网络获得相关分子的非共价结合数据，进而促进药物研发和构建结合预测模型。
4. Chemogenomics Knowledgebase, CGKB: 提供新的探索性计算工具/算法和化学库资源，以化学基因组学规模进行计算机辅助药物设计和发现。
5. SuperTarget Database: 该数据库集成了与医学指征、不良药物反应、药物代谢、途径和靶蛋白的 Gene Ontology (GO) 术语相关的药物信息。

6.8 多组学生物样本库

- 蛋白质-代谢通路数据库

1. KEGG 使用 UniProtKB+AC/ID 工具可将蛋白质 ID 编号转化为 KEGG 中的编号，然后使用 KEGGREST 包的 keggGet 函数即可查找对应的代谢通路。
2. Reactome 可以使用其 analysis 工具对 UniProt accession 蛋白质列表查询代谢通路。
3. HumanCyc 付费的
4. ConsensusPathDB 使用了和 KEGG 数据库相同的代谢通路数据。

- 蛋白质互作数据库

1. IntAct
2. DIP
3. MINT
4. HPRD
5. BioGRID

6. SPIKE
7. UniHI
8. 3DID
9. STRING
10. ConsensusPathDB