# 第 1 章　引　　言

## 1.1　研究背景与意义

人们对生命活动的理解总是伴随着各种生物技术的革新和发展而不断加深。从单个细胞，如受精卵，经过复杂的细胞分裂、分化发育等生命过程，进而演化成一个完整的个体，其背后的基因调控机理、分子运作机制等核心问题吸引着无数科学家锲而不舍地进行探究。英国科学家克里克在 1958 年提出“DNA→RNA→ 蛋白质”中心法则[1] 距今已有六十余载。现代生物学的研究让人们对基因调控机理等重要基础科学问题能够进行更加多层次、高精度的深度解读。21 世纪初期，人类基因组计划的完成让人们距离破解人类遗传密码迈出了一大步，为解析人类遗传密码提供了宝贵的基因组序列信息[2]。而在随后的研究中大家发现，在由大约 30 亿碱基对所组成的人类基因组中，真正能编码蛋白质的基因片段占比不到 1%。如何对占比 99%以上的非编码的基因组区域进行解读变得尤为重要，这对于深入理解基因调控机制、研究基因型与表型的关联关系、探索复杂遗传疾病的发生发展规律等具有重大的意义。

随着对基因组的研究不断深入，研究者们发现人类基因组中除几万个编码蛋白质的基因外，非编码区域的片段在基因转录等活动中也起了重要的调控作用，其中就包括如启动子（promoter）、增强子（enhancer）、沉默子（silencer）等重要的基因调控元件。这些调控元件往往不是单独地参与基因调控活动，而是多种调控元件形成复杂的调控网络与细胞中的蛋白质，如转录因子等一起共同参与基因调控[3]。在复杂的基因调控系统中，任何环节出现错误都可能经过一系列的连锁反应最终体现在个体表型的差异上[4]。不少疾病的产生与发展都与基因调控的失衡和错乱有着

直接或间接的关系[5]。遗传学研究的核心问题之一便是找到与疾病相关的基因与突变位点。这种突变位点如果在基因编码区域，则可以通过其转录得到的 RNA 或者编码得到的蛋白质来研究其突变对生命活动的影响。如果突变发生在非编码区域，那么对其机理的建模分析就变得极为困难。全基因组关联分析（GWAS）的研究告诉我们人体基因组上绝大多数的突变都存在于非编码区。因此，非编码区域的功能建模、突变与疾病表型的关系研究也显得前所未有的重要。

得益于以高通量测序技术为代表的生物技术突飞猛进的发展[6]，针对不同物种、不同器官、不同组织、不同细胞型的跨尺度、多模型的生物大数据得以大量积累。这为我们试图解析基因调控机理、破解遗传密码提供了宝贵的数据支持。如 ENCODE[7]、Roadmap[8] 等公共数据库提供了诸如基因表达、染色质开放性、转录因子绑定位点、DNA 甲基化等数据。这些涵盖了基因组、转录组、蛋白组、表观基因组不同层次的丰富数据能让我们深入理解基因调控的过程、解析基因调控的机理。这些大规模、跨尺度、多模态的量化生物大数据为研究基因调控机理提供了宝贵的数据基础。

另外，单细胞技术的出现让人们对细胞调控的理解进入全新的层次——单细胞水平的基因调控[9]。上述基因表达与染色质开放性等生物信号已经能在单个细胞上得到测量。单细胞测序很大程度上解决了传统细胞群测序技术难以处理的细胞异质性、细胞间时空相互作用、细胞特异性分化等重要问题，这使得单细胞测序技术很快便成为现代生命科学研究的基本手段与工具之一。对这些细胞群水平和单细胞水平的跨尺度、多模态生物大数据解读已经成为生命科学的核心内容之一。

人们注意到上述生物大数据目前正以几何级数的速度进行快速生产和积累，这些数据的分析亟需一些表现力强、通用性高的计算模型[10]。一方面，计算模型可以利用现有数据对基因调控的过程进行精准建模，如针对单种调控元件的活性与其所处的基因组环境进行关系型建模，亦或是对调控元件、转录因子、基因等多种参与基因调控的基础元件进行调控网络的构建[11]，此类计算模型能让人们对基因调控的机理在一定前提假设下进行精确的刻画和描述，让人们对基因调控这一复杂的生物过程有一个更为直观的认识。另一方面，以机器学习方法为代表的计算模型在生物

大数据的大量应用让人们意识到，合适的机器学习模型能对大量生物数据的特征进行良好的学习，并对未知的生物信号进行有效的预测和判断。尤其是在计算机视觉、自然语言处理等多个领域已经取得突破性进展的深度学习技术[12]，已经被证明为对大规模数据进行信息挖掘的有效工具。

这些强有力的计算模型不仅能够帮助有效解析大规模生物数据中所蕴含的生物知识，同时也能克服目前生物大数据自身存在的局限性。如仅仅依靠大规模的生物实验数据难以涵盖所有的物种、器官、组织、细胞系环境。而强有力的机器学习方法便可以在一定程度上弥补生物实验成本过高、部分生物数据缺失等问题。计算生物学家针对不同生物大数据的分析和计算已经提出了大量的计算生物模型。在这一点上，计算生物学模型的发展与大规模的生物实验数据的产生和积累是相辅相成的。生物大数据加上解析生物大数据的高效计算模型成为人们探索现代生命科学奥妙、研究基因遗传密码的两个不可或缺的重要因素。

本书正是考虑到生物大数据在当前正以前所未有的规模高速进行生产和积累，这对生物计算模型的计算性能、效率等提出了越来越高的要求。要想全方位地解读这些生物大数据，目前还存在着对生物大数据所蕴含的生物功能预测不够准确、对生物大数据的多源异质性的协同分析不够细致等局限。本书以重要的表观遗传学信号染色质开放性为研究主线，以机器学习方法尤其是深度学习方法为主要手段，针对以染色质开放性为代表的生物实验数据进行了系统性的解读、建模和分析。本书回答的科学问题包括：细胞群染色质开放性预测中进化保守性信息的整合问题、细胞群染色质开放性预测中基因组短片段词频特征的结合问题、细胞群染色质开放性预测中先验生物知识融合问题、基于染色质开放性数据的遗传学数据解读问题、单细胞染色质开放性数据中细胞类型发现问题、单细胞染色质开放性数据与单细胞基因表达数据协同分析问题等。

### 1.1.1 高通量测序技术

迄今为止，DNA 测序技术已经经过了主要三代技术的迭代。Sanger 等首先在 1977 年完成了噬菌体的超过 5000 bp 的基因组测序[13]。第一代测序技术就是在 Sanger 等所提出的基础上改进而来的。最著名的应用便是人类基因组计划[2,14]。第一代测序技术具有长读段、高成本、低通量等

特点，这会严重影响测序的效率及 de novo 测序、转录组测序等新型测序技术的普及。

第二代测序技术则主要基于 Roche 公司的 454 技术[15]、Illumina 公司的 Solexa 技术、Hiseq 技术[16-17] 等。相比于第一代测序技术，第二代测序技术能同时测量百万级别的核酸分子序列，并且让测序成本进一步降低且测序通量、测序多样性进一步增加，同时测序读段相比于第一代要小很多。第二代测序技术如今已经极为广泛地应用在基因组测序、转录组测序、蛋白组测序、表观遗传组测序上[18-21]，从而极大地丰富了序列技术的应用场景。时至今日，第二代测序技术仍然是使用最为广泛的测序技术。

第三代测序技术则是以 PacBio 公司的 SMRT 技术及 Oxford Nanopore 纳米孔技术[22] 为代表，其最大的特点便是不需要 PCR 分子扩增，可直接对单个 DNA 或 RNA 分子进行测序，具有很高的便携性。目前该技术正处于快速发展阶段，距离其大规模的广泛应用还需要一定时间。如图 1.1 所示，随着测序技术的不断迭代和更新，人体基因测序的成本已经得到了显著的下降。

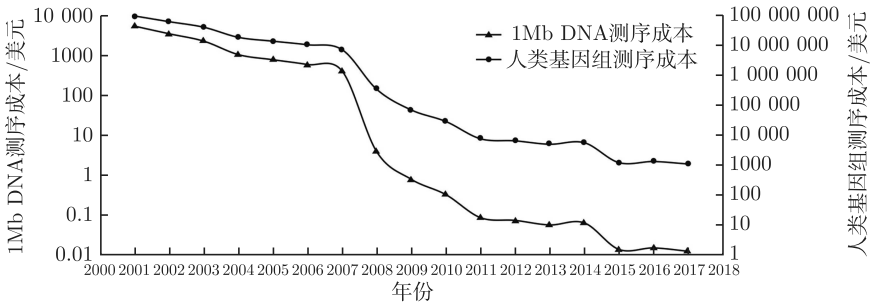


**图 1.1　基因测序成本随年份的变化[23]**

如果从测序样本分辨率的角度来看待测序技术的革新与演变，那么测序技术则经历了从细胞群测序（bulk sequencing）到单细胞测序（single-cell sequencing）的发展过程。传统的细胞群测序技术往往需要百万甚至更多的细胞组成测序样本，得到的测序数据体现的是细胞群样本的信号强弱。而单细胞测序技术的出现为揭示复杂的生物活动，诸如细胞分化发育等，提供了重要的线索[24-26]。单细胞技术从很大程度上解决了传统细

胞群测序技术无法解决或很难解决的稀有细胞类型识别、人类细胞图谱构建等重要科学问题。

剑桥大学汤富酬等在 2009 年率先将单细胞高通量测序技术应用在人类卵细胞的单细胞基因组分析上[27]。在此之后，单细胞测序技术迅速发展壮大并扩展到基因组、转录组、表观遗传组、三维结构基因组等类型的组学数据中[28-32]，从而极大地丰富了单细胞测序技术的应用范围。同时以分子液滴[33]（droplet）、微孔芯片[34]（microwell）及组合索引技术[35]（combinatorial indexing）为代表的单细胞建库技术的出现，让单细胞测序实验的通量直接提升至数千细胞的量级，使得大规模的单细胞数据研究成为可能。图 1.2 展示了上述三种主流细胞标记与建库方法的特性。单细胞技术的出现直接提高了测序数据的分辨率——从传统几十万至数百万的细胞直接提升到了单个细胞的测量精度，极大地丰富了生物学的研究手段，使得人们对细胞中各类生物大分子的活动特征描述更为准确，从而能够更加清晰地理解组织、器官与生物体的发育和演化过程，并为解析生命系统中细胞分化发育规律、研究疾病的产生机理的重要科学问题带来了新的契机。

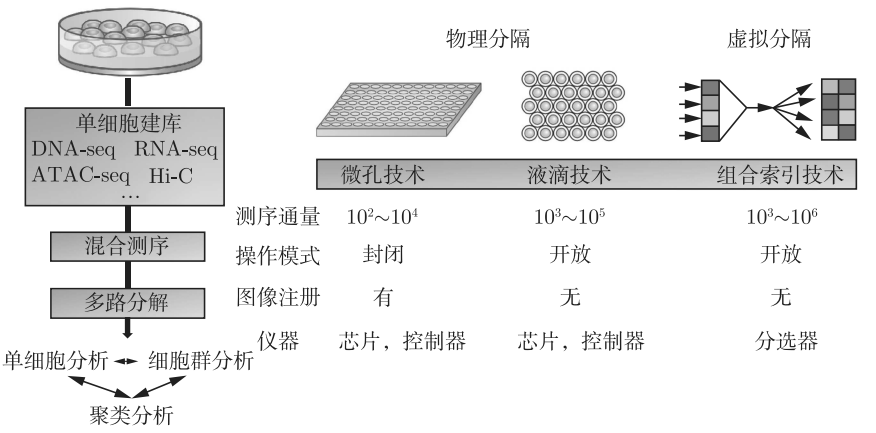


**图 1.2　高通量单细胞测序主流细胞标记与建库的方法**

### 1.1.2　染色质开放性

在了解染色质开放性之前，必须了解真核细胞内染色质的基本结构。尽管人体基因组 DNA 由大约 30 亿对碱基组成，DNA 分子拉直后的物

理长度大约为 2 m，但 DNA 并不是规则地线性排列在细胞核中，而是与核小体折叠缠绕成更为紧致的形态。真核细胞的染色质往往会紧密地包裹在一系列的核小体中。每个核小体被长度大约为 147 bp 的 DNA 片段折叠环绕两圈[36-38]，核小体及缠绕在其上的 DNA 片段便构成了染色质的基本组成单元。核小体的核心由四种不同的组蛋白组成，可以通过共价修饰过程改变自身的结构[39-40]。核小体在整个基因组的定位具有重要的调节功能，并且能够直接影响转录因子绑定位点的分布，从而进一步影响 DNA 参与的生命活动，如基因转录、DNA 修复、DNA 复制等[41]。

染色质的开放性则可以定义为 DNA 在与核小体结合后再与其他生物分子（比如转录因子）结合的能力。如图 1.3 所示，在染色质闭合区域，核小体分布紧密，DNA 被核小体覆盖程度高；而在染色质开放区域，核小体分布较为稀疏，部分“裸露”的 DNA 则相对更容易地与转录因子（TF）等蛋白质相结合，从而进行基因转录等生命活动，所对应的“裸露”的 DNA 区域则相对而言更加“开放”。染色质开放性具有很强的动态性，这种动态性在细胞特异性与时间空间特异性上均有体现，因此，染色质开放性也体现了不同环境、不同时刻的细胞功能基因组的状态。

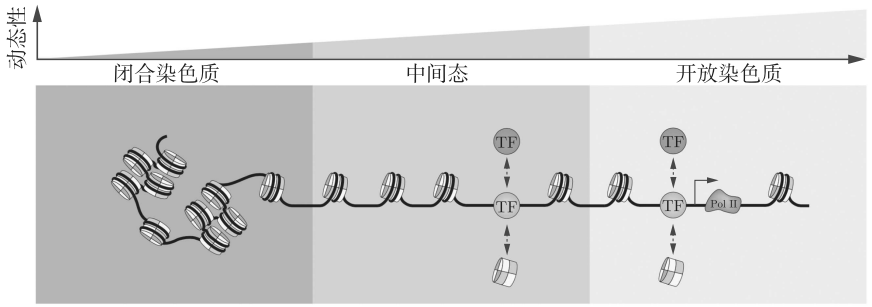


**图 1.3　闭合染色质与开放染色质示意图[42]**

染色质开放性的区域在整个基因组中占比仅为 2%∼3%，但是却捕获了超过 90%的转录因子结合区域，仅有为数不多的几个转录因子与 DNA 结合的区域集中分布在异染色质区域[43]。一方面，转录因子与组蛋白及其他的染色质绑定蛋白通过竞争调节核小体在染色质上的分布[44-45]；另一方面，不同细胞系的染色质开放性也意味着不同的转录因子绑定的形式[46-48]。转录因子具有非常广泛的功能作用，能为基因转录活动提供一

定的动态调节，并建立维持共同基因组在不同细胞类型下进行基因转录活动的表观遗传通道。因此，染色质开放性不仅反映了转录因子的绑定结合能力，也反映了在该基因区域的调控潜能。

目前主流的染色质开放性实验测量方法通常通过量化染色质对酶促甲基化或 DNA 分子裂解的敏感度来实现。原则上来讲，染色质开放性应该取决于与 DNA 分子片段进行绑定交互的蛋白质分子类型。然而有研究发现，染色质开放性对不同的蛋白分子类型具有高度的保守性[49]。1973 年，Hewish 等率先使用 DNA 核酸内切酶将染色质片段化，表明核小体在整个基因组上具有周期性超敏反应[50]。具体体现在整个基因组的 DNase 超敏性位点之间呈现出 100∼200 bp 的周期性。而在 1958 年 PCR 技术被引入后[51]，人们能够利用核酸内切酶与连接介导 PCR（ligation-mediated PCR）对染色质的特定片段的开放性进行定量测量[52-53]。随着高通量测序技术的出现与发展，现代生物学中对染色质开放性的测量已经几乎完全被高通量测序技术方法所取代。DNase-seq 技术便是利用非特异性的核酸内切酶对 DNase I 超敏位点（DHS）进行全基因组尺度的切割[54-55]。全基因组尺度的染色质开放性数据显示在启动子和转录起始位点（TSS）的近端仅发现了少数 DHS 单位点，超过 80%的染色质开放区域都位于远端的增强子区域。DNase-seq 实验流程如图 1.4 左侧所示，DNA 片段会经历 DNA 切割、黏性末端融合、建库、测序等流程。

另一种被广泛应用于测量全基因组染色质开放性的技术为 ATAC-seq，在 2013 年由斯坦福大学的 William J.Greanleaf 与 Howard Y.Chang 等提出[49]。其基本原理是利用高活性的 Tn5 转座酶将全基因组的染色质开放区域进行切割并添加测序适配接头（adaptor），这些测序适配接头里具有已知的 DNA 序列标签，从而可以利用这些已知的 DNA 序列标签进行建库、PCR 扩增等。其实验流程如图 1.4 右侧所示。与 DNase-seq 技术相比，ATAC-seq 技术具有可操作性强、对细胞数量要求低（几百）、实验重复性好等优点。ATAC-seq 技术已经逐渐成为测量全基因组染色质开放性的首选方法[56]。

除应用广泛的 DNase-seq 及 ATAC-seq 技术外，基于微球菌核酸酶切割核小体的 MNase-seq 技术[57] 及利用甲醛、酚氯仿抽提分离从而获取裸露 DNA 的 FAIRE-seq 技术[58] 也得到了一定范围的应用。但这些

技术和 DNase-seq 具有的共同缺点就是对细胞数量要求过高，从而在一定程度上限制了这些技术的应用场景和范围。具体而言，MNase-seq 技术测量的目标区域为染色质中的核小体区域，其实验过程需要保证在建库过程中精准控制酶量，对实验技术要求较高。FAIRE-seq 技术在细胞数量上的要求相对于 MNase-seq 与 DNase-seq 略低，但其实验数据存在信噪比低等问题，这对解读实验数据与其下游分析造成了一定的困难。整体而言，ATAC-seq 技术存在细胞需求量小、实验可操作性强、测量准确度高等优点，这些特点也使得 ATAC-seq 技术的应用越来越广泛。DNA 元件百科全书数据库 ENCODE[7] 便收录了不同实验技术所测量的染色质开放性数据，但仅有 DNase-seq 与 ATAC-seq 数据包含的较为全面。表 1.1 展示了上述四种染色质开放性的实验获取方法的不同特点与比较。

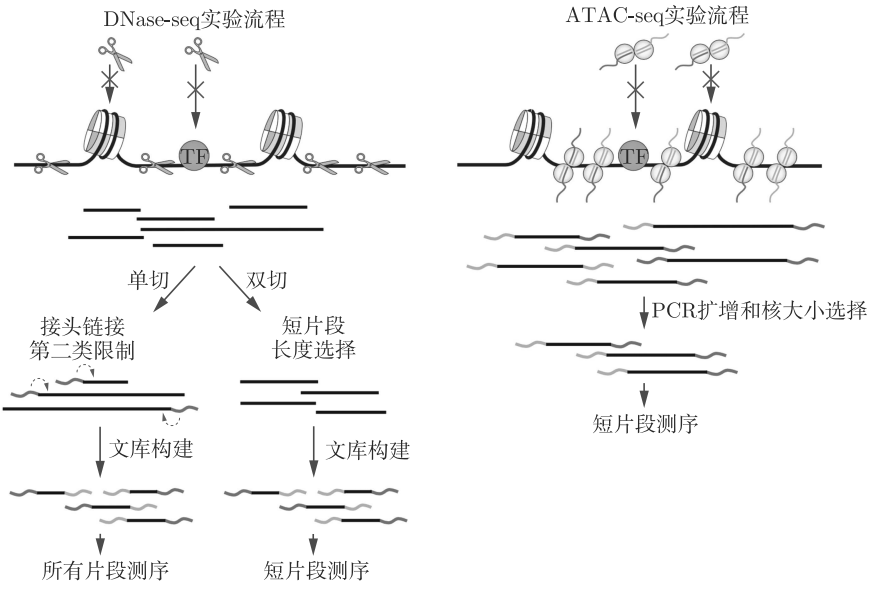


**图 1.4 DNase-seq 与 ATAC-seq 实验流程[42]**

随着单细胞测序技术的出现与发展，研究者们也尝试将单细胞测序技术用于染色质开放性的测量中，即获取单个细胞的染色质开放性状态，由于二倍体生物，如人类，其遗传信息的拷贝数目理论上最大即为 2。意味着单细胞染色质开放性的测序方法需要捕获单个细胞中的单个或者拷贝

数为 2 个的 DNA 片段。基于上述细胞群染色质开放性测序技术 ATAC-seq 的成功经验，单细胞染色质开放性需要解决的最大问题便是细胞分离和建库的技术。目前主流细胞分离与建库技术有两种，即组合索引技术[59]和微流体技术[60]。

**表 1.1 获取染色质开放性的不同实验方法**

| 方法名称 | 细胞数目 | 建库原理 | 目标区域 | 优缺点 |
|---|---|---|---|---|
| MNase-seq | $10^7$ | 微球菌核酸酶 MNase 消化染色质上未被核小体或蛋白质保护的 DNA | 核小体区域 | 需要大量细胞，精准控制酶量 |
| FAIRE-seq | $10^5 \sim 10^7$ | 利用甲醛固定酚氯仿抽提分离获取裸露的 DNA | 染色质开放区域 | 信噪比低，数据解读困难 |
| DNase-seq | $10^7$ | 利用 DNase I 优先切割核小体被取代的 DNA 序列 | 染色质开放区域，侧重于转录因子结合位点 | 需要大量细胞，样本制备复杂 |
| ATAC-seq | 500 ∼ 50 000 | Tn5 转座酶切割并插入未经核小体或者蛋白保护的 DNA | 染色质开放区域 | 细胞需求量减小，可操作性强，线粒体数据可能污染 |

组合索引技术提供了一种非常精巧的策略，从而能对数千个单细胞 ATAC-seq 的库文件进行条形码编码[59]，在这种方法中，使用唯一条形码编码的 Tn5 转座酶在纯化的细胞核上进行多次转座反应，共享相同转座反应的一对细胞则在随后合并和分裂操作期间进行共同分离，随后被分别筛选进入含有第二轮条形码编码的 PCR 引物的多孔板中。基于此特点，组合索引技术不需要单独分离单个细胞即可对成千上万的单细胞进行建库。其流程如图 1.5 所示。

基于微流体技术的单细胞 ATAC-seq 技术（scATAC-seq）同样由原 ATAC-seq 技术的发明者 William J.Greanleaf 与 Howard Y.Chang 等提出[60]，微流体技术能够保证一个液滴里仅包含一个细胞核，从而方便完成细胞分离任务，其次便是 Tn5 转座酶的酶切作用、酶切短片进行 PCR 扩增等过程，最终去掉油滴，对序列进行测序操作。其操作流程如图 1.6 所

示[60]。尽管每次实验能处理的细胞数量不如组合索引技术，但细胞索引技术的建库复杂度要高出不少。目前 scATAC-seq 技术已被 10X Genomics 等公司在基于液滴的微流体一站式平台上实现并商用。scATAC-seq 技术也逐渐成为单细胞染色质开放性测量的主流方法。

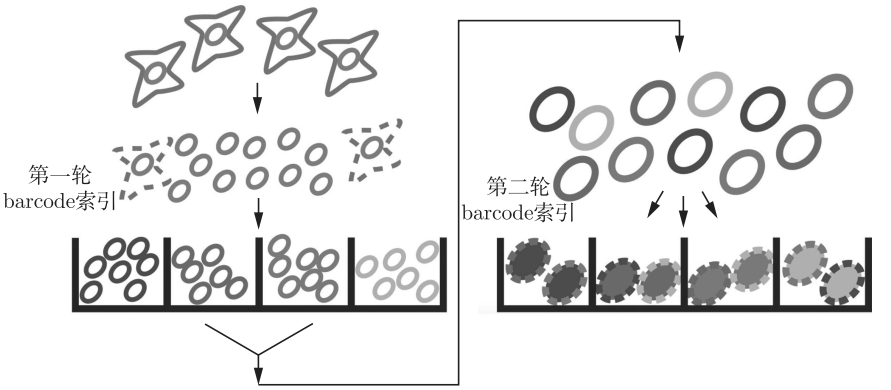


**图 1.5　组合索引技术流程示意图[59]**

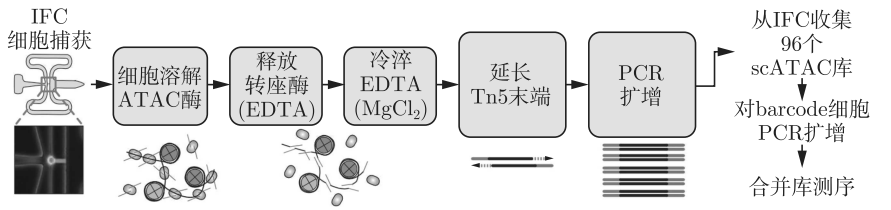


**图 1.6　基于微流体技术的 scATAC-seq 流程示意图[60]**

### 1.1.3　基因调控机制

细胞通过染色质的复杂结构承载大量遗传及调控信息，即 DNA 包裹在组蛋白周围并与之紧密包装结合在一起。为了在特定的编码区表达基因，染色质将打开，DNA 可通过与增强子、启动子等细胞调控元件相互作用而形成基因调控的核心组成部分。此外，细胞的内聚复合物、序列特异性转录因子和 RNA 聚合酶 Ⅱ 等被召集起来共同以精细的方式调节基因表达水平[61]。人体内长度大约为 2 m 的 DNA 分子通过紧密而复杂的折叠存在于细胞核中，这个过程往往会牵涉 DNA 对组蛋白的包裹及核小体结构的形成。研究者们通常认为染色质的结构可以影响基因表达、