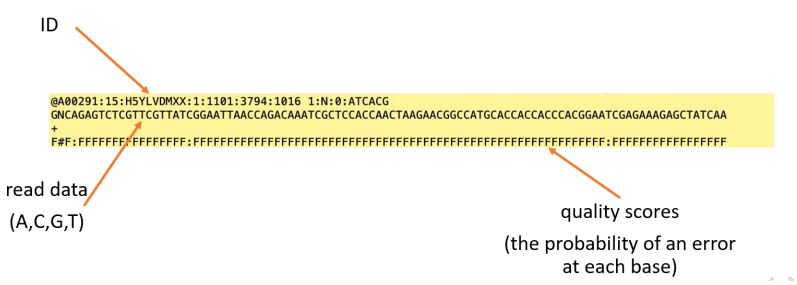
GP

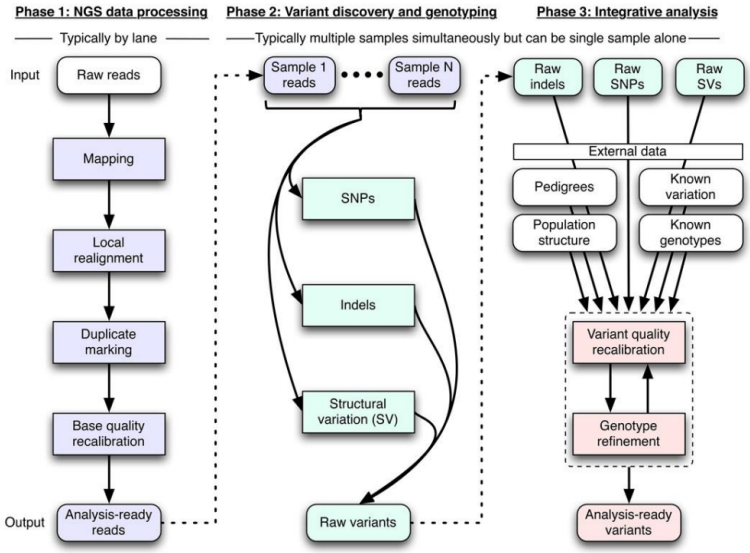
- NGS (next-generation sequence)
  - 合成测序(**Sequencing by synthesis**)
    - 准备DNA**(Prepare genomic DNA)**
      - 随机片段化DNA，并给片段加上adatpers
    - 固定DNA**(Attach DNA to surface)**
      - 在流式细胞通道的内表面随机固定单链片段。
    - 桥式扩增**(Bridge amplification)**
      - 加入无标记的核苷酸和酶扩增片段
    - 解旋**(Denature the double stranded molecules)**
    - 依照模板合成DNA，同时放出荧光信号，不同核苷酸对应不同信号
  - 主要数据分析步骤
    - 质控 Quality control
      - FASTQ format



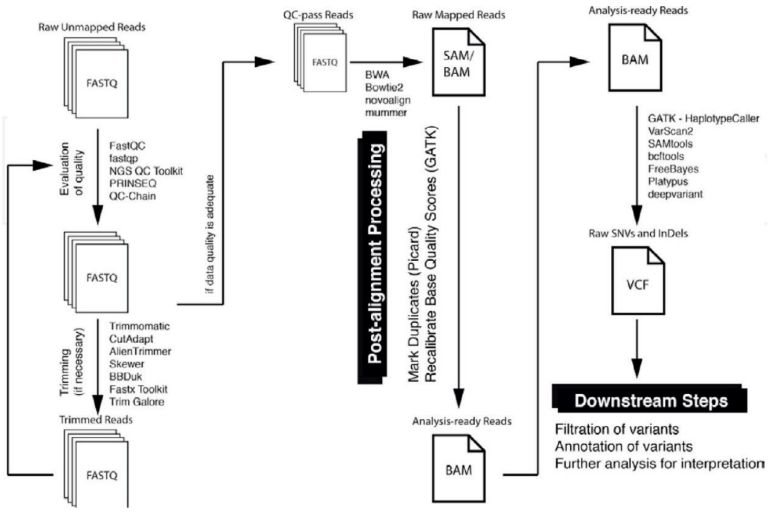
    - 比对 alignement
      - 为什么BLAST算法不合适?
        - 1. 速度太慢
        - 2.
        - 3.
      - BWT (Burrows–Wheeler_transform)算法
        - 运行逻辑：
          - 将输入字符串结尾加上$并向右轮换，将轮换后所得的矩阵以行为单位以ASC II 码为标准，上下排序，输出矩阵的最后一列与?
        - 优点：
          - 1. 通过输出的字符串+sorted这个字符串，可以快速搜索相应字符内容
          - 2. 尽可能将相同字符串放在一起，提高压缩效率
      - Quality control Alignment
        - 50% of reads mapped/unmapped would be thought as fail.

        - Transcript assembly/quantification
            - 样本之间的差异count显示表达差异
            - individual reads可以被mapped到不同的transcripts或isoforms上
        - Peak-calling
            - 寻找reads富集的区域--peaks
    - NGS的应用
        - *de novo* genome assembly
            - 例如：Using next-generation sequencing technology alone, we have successfully generated and assembled a draft sequence of the giant panda genome
        - Genome resequencing
            - 对于特殊个体，检查是否有variant gene
        - Clinical applications
            - 检查是否携带variant gene
        - Sequencers as counting devices
- WGS (Whole genome sequence)
    - history of WGS
    - De novo assembly vs Variant calling
        - De nove assembly
            - 物种的基因组是未知的
            - 从头开始构建基因组
        - Variant calling
            - reference genome是已知的
            - 目的是了解个体基因组的variances
    - De nove assembly
    - Variant Calling
        - 种类 types
            - 单核苷酸畸变 Single Nucleotide Aberrations
                - 单核苷酸多态性 Single Nucleotide Polymorphisms (SNP)
                    - 可以发生在一个物种的所有成员中
                    - 在群体中的某个特定位点以一定频率出现
                    - 在群体中已经确认过
                    - 收录于dbSNP(http://www.ncbi.nlm.nih.gov/snp)
                - 单核苷酸变异 Single Nucleotide Variations (SNV)
                    - 畸变只出现在一个人身上
                    - 发生频率低，不常见

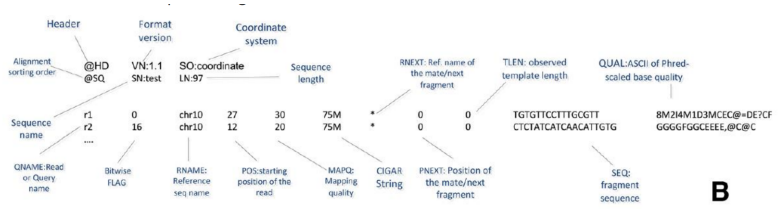
        - 在群体中不常见
    - 短插入或短删除 Short Insertions or Deletions (indels)
    - 结构性变化 structure variants (SVs)
        - large deletion or insertion
        - chromosome rearrangement
- 工作流程 workflow
    - phase 1 : NGS data processing
    - phase 2 : variant discovery and genotyping
    - phase 3 : integrative analysis
    - overall workflow



- Pipeline
    - phase 1 : Mapping
        - 使用mapping算法将reads在初次比对到reference genome上
        - 修补初始比对结果
            - 局部indels重新比对
            - 消除重复的分子
        - 生成technology-independent的SAM/BAM文件
    - phase 2 : Discovery of raw variants
        - 使用SAM/BAM文件，分析所有含有SNPs, SNVs, short indels, SVs的位点
    - phase 3 : Discovery of analysis-ready variants
        - 从第二步获得的raw variant中分离真的多态性位点
    - overall pipeline

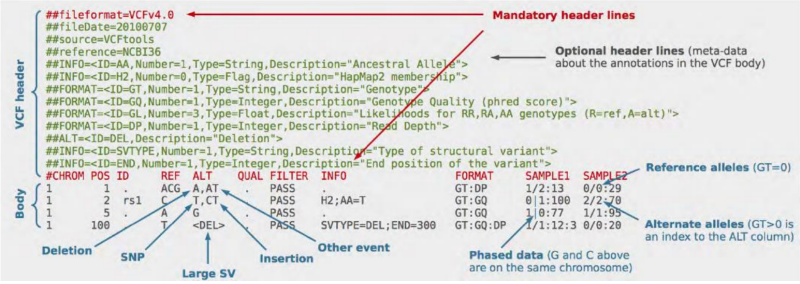


- File format
  - SAM/BAM
    - BAM文件是SAM的二进制压缩文件
    - 内容
      - Header
        - 包含文件的所有信息
          - 样本信息
          - 样本长度
          - 比对方法
      - Alignment
        - 包含
          - read的名称
            - 染色体
            - 起始位点
            - 比对质量
            - 匹配描述字符串
          - 序列
          - 质量、
          - 比对信息
          - custom tag
      -

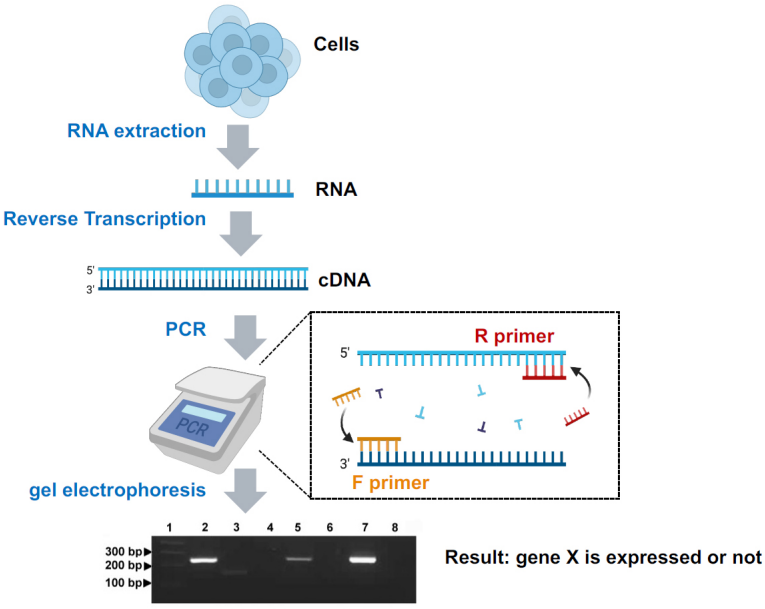


  - VCF

        - Header
            - VCF的版本
            - variant caller的版本
        - Data Line
            - 每行包含单个variant的信息

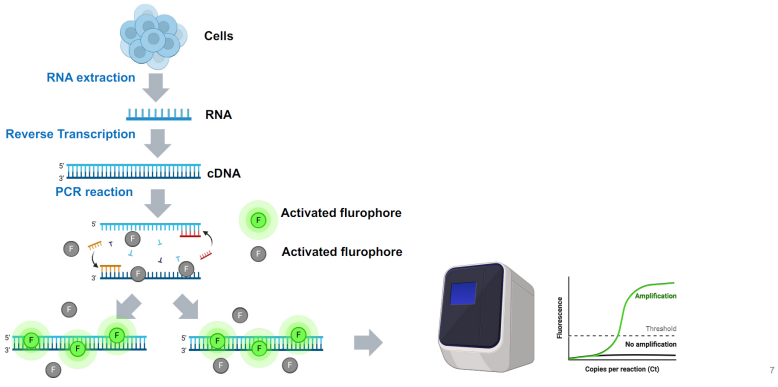


- 应用 application
    - 研究生物进化 Evolution studies
        - 发现新的病原体 如：2019 nCov
    - 微生物分析 Microbiome analysis
        - 食品安全 food safety
        - 健康有关的传染病 Health-related infection
        - 疾病相关性 Disease association
    - 群体特征 Population traits
    - 疾病关联性研究 Disease association study
    - 个性化用药 Personalized medicine
    - 系统发生学 Phylogenetics
        - Phylogenetic tree 构建系统发生学树
        - pangenome analysis 确认核心基因组及附属基因
        - effective population size 预测数量
        - transmission map 确认传染地图
        - recombination hotspots 预测多样化选择下的regions
        - genome-wide association studies 确认SNP等相关variants
- 单分子测序 Single molecular sequencing
    - PacBio
        - https://www.zhihu.com/topic/20015076/top-answers
    - Nanopore
- Gene Expression Analysis
    - technique
        - 1th generation: RT (Reverse Transcription)-PCR 反转录PCR
            - 流程：



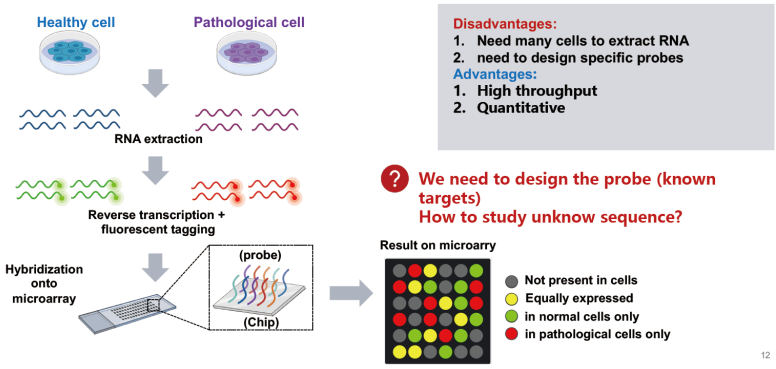
- Disadvantages:
  - 1. Single gene at a time 通量低
  - 2. Need many cells to extract RNA
  - 3. need to design specific primers
  - 4. result not quantitative 无法定量
  - 5. low sensitivity 敏感度较低
- Advantages:
  - 1. Fast
  - 2. Cheap

- 2nd generation technique: RT-qPCR (quantitative) 可定量反转录PCR
  - 流程:



  - Disadvantages:
    - 1. Single gene at a time
    - 2. Need many cells to extract RNA
    - 3. need to design specific primers
  - Advantages:
    - 1. Fast

    - 2. Cheap
    - 3. Quantitative
- 3rd generation technique: microarray 微矩阵
    - 流程：



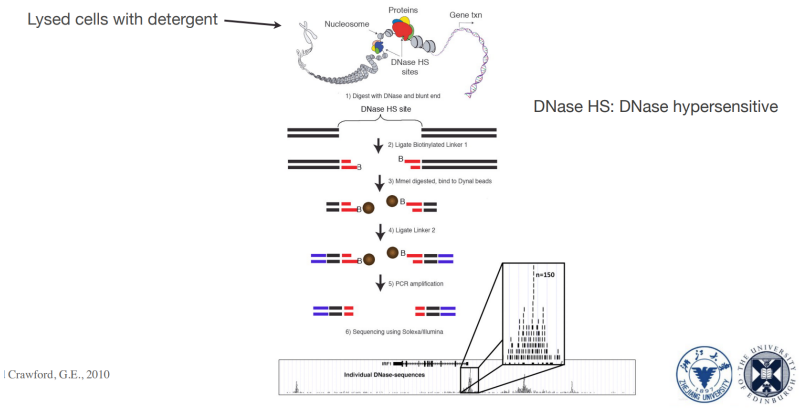
    - Disadvantages:
        - 1. Need many cells to extract RNA
        - 2. need to design specific probes
    - Advantages:
        - 1. High throughput
        - 2. Quantitative
- 4th generation technique: Next Generation Sequencing based RNAseq
    - 流程：
        - RNA转录为cDNA
        - 将cDNA裂解为片段并加上adapter
        - 在流式细胞仪中桥式扩增片段
        - 是基于样本合成测序
        - 比对，分析
    - Disadvantages:
        - 1. Need many cells to extract RNA
        - 2. expensive
        - 3. data analysis complicated
    - Advantages:
        - 1. High throughput
        - 2. Quantitative
        - 3. no prior information needed
- 5th generation technique:
    - single cell RNA-seq (scRNA-seq)
        - 流程

- 获得组织切片
- 使用FACS分离单细胞
- 构建RNAseq库并测序
- Disadvantages:
- 1. expensive
- 2. data analysis complicated
- 3. low coverage compared to NGS
- Advantages:
- 1. High throughput
- 2. Quantitative
- 3. no prior information needed
- 4. single cell level
- long read sequencing （Nanopore sequencing）
- 流程
- 将长read通过纳米孔
- 通过电流变化记录RNA序列
- Disadvantages:
- 1. Need many cells to extract RNA
- 2. expensive
- 3. data analysis complicated
- 4. high error rate
- Advantages:
- 1. High throughput
- 2. Quantitative
- 3. no prior information needed
- 4. long reads (>kbs): higher resolution
- Nature of mRNA-seq library RNA文库的性质
- Poly A enriched
- 只连接RNA片段
- cDNA
- 不含有内含子片段
- RNA seq Analysis
- FASTQC质控
- 工具
- Fastqc ——获得质控报告

        - Trimmomatic——trim 片段提高质量
        - Cutadapter——去除adapter
- Splice-aware alignment togenome 比对
    - 工具
        - Short reads aligner: BWA, Bowtie, Bowtie2 ...
        - De novo splice aligner: HISAT,HISAT2 …
        - De novo splice aligners thatalso use annotation optionally : Tophat, Tophat2, STAR …
        - Recommend： STAR, Tophat, HISAT
    - 以STAR为例
        - 步骤
            - Step1 seed search
            - Step2 deal with mismatch
            - Step3 deal with unmapped reads
            - Step4 Clustering, Stitching,Scoring
    - Assign counts to genes/isoform 将比对结果联系至基因
        - 工具
            - FeatureCount
            - HTSeq-count
    - Normalization 归一化结果
        - Why we need normalization?
            - The more you sequenced (total reads sequenced), the more reads you would get over each genes.
            - The longer the gene is, the more reads you would get over the genes.
        - RPKM (Reads Per Kilobases per Millionreads)
    - Differential expressed genes calling 寻找差异性表达的基因
        - 工具
            - DEseq2
            - EgeR
        - multiple testing correction 多重检验纠正
            - Bonferroni
            - FDR/Benjamini-Hochberg (most commonlyused)
                - 通常值为 1.5，2，4
            - Q-value/Storey method
                - 通常值为0.05

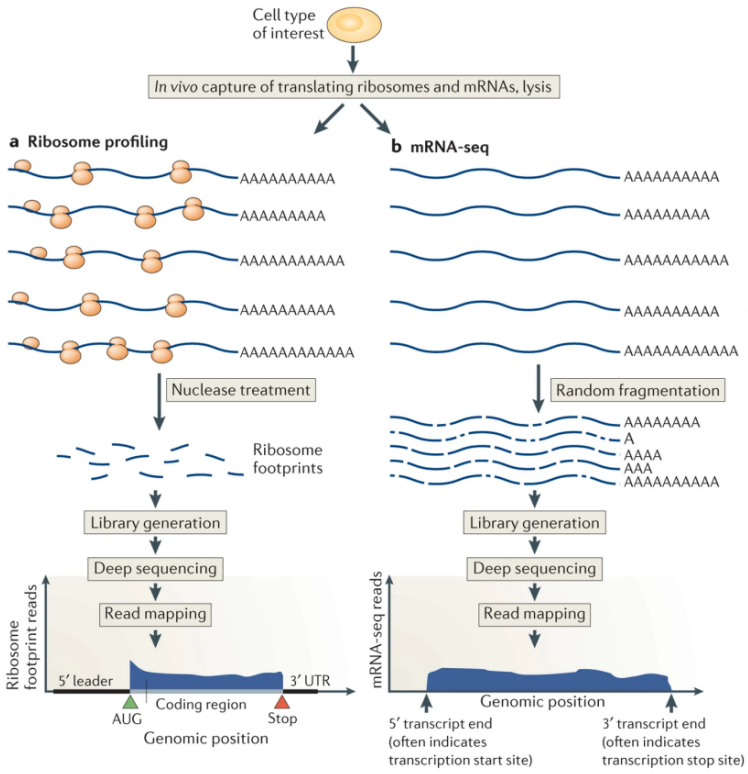
- Application 应用
  - 基因表达谱Gene expression profiling
  - 丰度估计Abundance estimation
    - 可变剪接alternative splicing
    - RNA编辑RNA-editing
    - 识别新的转录本novel transcript identification
  - 融合转录本fusion transcripts
  - 等位基因特异表达 Allele specific expression
    - prevalence of transcribed SNPs
- 例子 example
  - The identification of COVID-19 using NGS

- ChIP-seq and Other SequencingApplications
  - Adaptation
    - DNase-seq

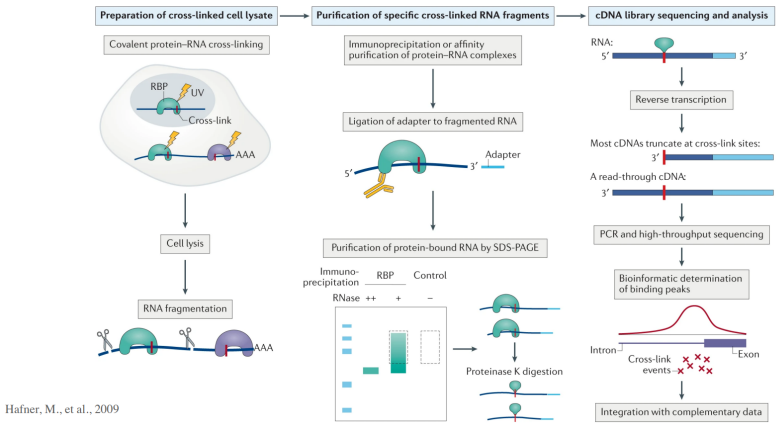


    - 用的DNase I内切酶识别开放染色质区域
  - ATAC-seq -> Assay for Transposase-Accessible Chromatin
    - 通过**转座酶Tn5容易结合在开放染色质的特性**，然后对Tn5酶捕获到的DNA序列进行测序。
    - 特点
      - 无需使用Tags或者antibody
    - 分析
      - QC，read trimming
      - Alignment
      - Peak Calling
        - To identify TF binding sites 寻找TF的结合位点
        - peak的定义
          - Count based 较多reads数量落在同一个region时，可称之为peak

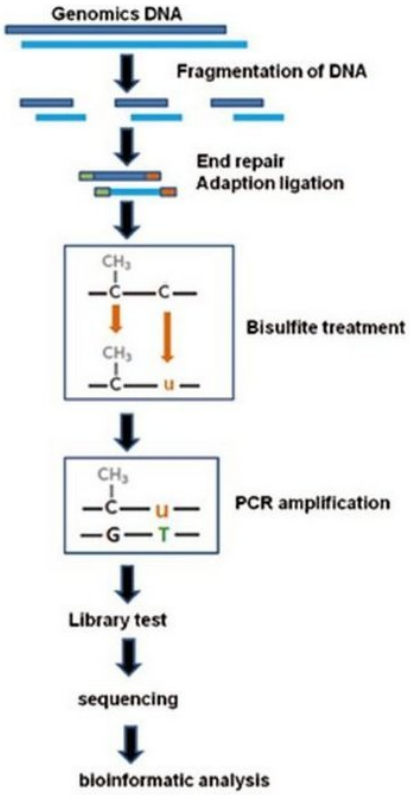
        - Shape based 候选结合位点。当read的分布符合预期分布时，任其为一个peak
    - QC，data visualisation
- FAIRE-seq
    - 先进行超声裂解，然后用酚-氯仿富集
- Ribosome Profiling
    - 用于蛋白质合成的定量分析，通过识别正在进行主动翻译的转录本，提供某一时刻的细胞内蛋白质情况
    - 应用
        - Investigate translational control and measure gene expression 研究基因表达
        - Identify translation start sites 确认翻译起始位点
        - Determine the rate of protein synthesis 确定蛋白质合成速率
        - Predict protein abundance 预测蛋白质丰度
    - 流程



- CLIP-seq -> Cross-linking and Immunoprecipitation 交联免疫共沉淀
    - 利用RNA和RNA结合蛋白在UV下可以结合的原理，沉淀RNA
    - 流程

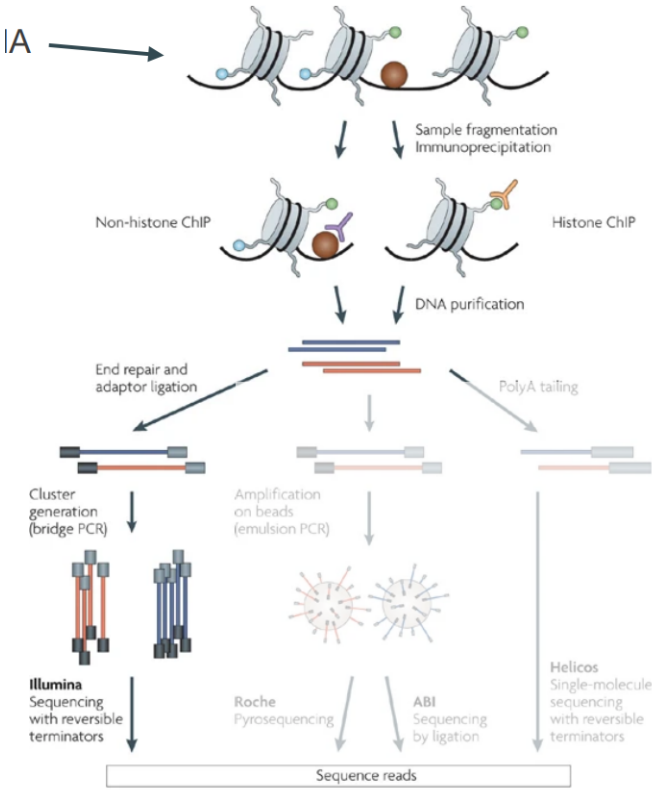


- Bisulfite Sequencing 亚硫酸氢盐测序
  - 可以将未甲基化的C碱基变为U，通过PCR变成T，使原本甲基化修饰的C被区分开来
  - 流程



    - 基因组DNA被打断为100-500bp的片段
    - DNA片段末端修复，3端连接A碱基，连接测序头
    - Bisulfite 处理
    - 脱盐处理，PCR扩增后文库片段大小选择
- MeD-seq with LpnPI <https://genome.cshlp.org/content/28/1/88.long>
  - LpnPI
    - a DNA methylation-dependent enzyme that is blocked by a fragment size smaller than 32 bp 一种DNA甲基化依赖酶
    - 防止甲基化密集DNA的完全切割
  - 优点

        - 测序深度小，小于亚硫酸氢盐测序的十分之一
    - 应用
        - 病人特异性差异化甲基片段 Patient-specific differentially methylated regions（DMRs）
        - 器官特异性 DMRs
        - 非活性X染色体启动子甲基化的差异
- ChIP-seq



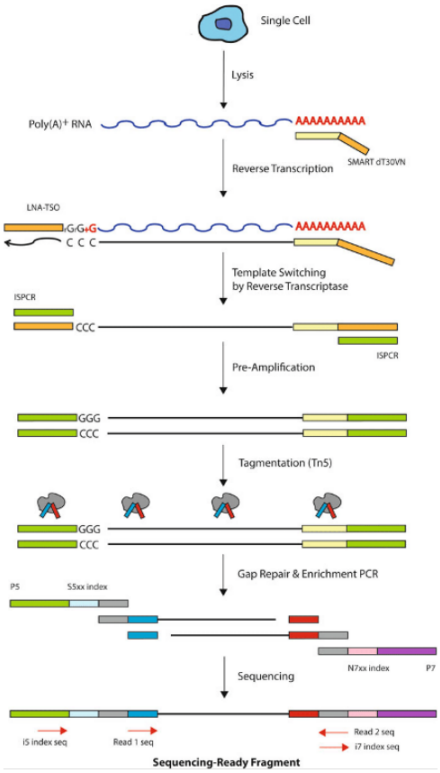
    - 流程
        -
    - map protein-DNA interaction place 可以确认的位点
        - Transcription factor binding sites 转录因子结合位点
        - DNA-binding proteins (HP1a, Lamins, HMGA, ...) DNA蛋白结合位点
        - RNA Pol II occupancy
        - Histone modifications 组蛋白修饰位点
    - Technical observations
        - 抗体的选择会影响结果
        - cross-linking交联时间在2-30min之内
            - 过度的交联会影响抗体可及性以及sonication的效率
        - Sonication的结果应控制DNA片段大小在200-1000bp之内
            - sonication时间过长会影响核小体和DNA的连接
        - 测序效率与碱基组成有关

        - 通常需要重复1-2次
    - histone code

Type of modification	Histone							
	H3K4	H3K9	H3K14	H3K27	H3K79	H3K122	H4K20	H2BK5
mono-methylation	activation[8]	activation[9]		activation[9]	activation[9][10]		activation[9]	activation[9]
di-methylation		repression[4]		repression[4]	activation[10]			
tri-methylation	activation[11]	repression[9]		repression[9]	activation,[10] repression[9]			repression[4]
acetylation		activation[11]	activation[11]	activation[12]		activation[13]		

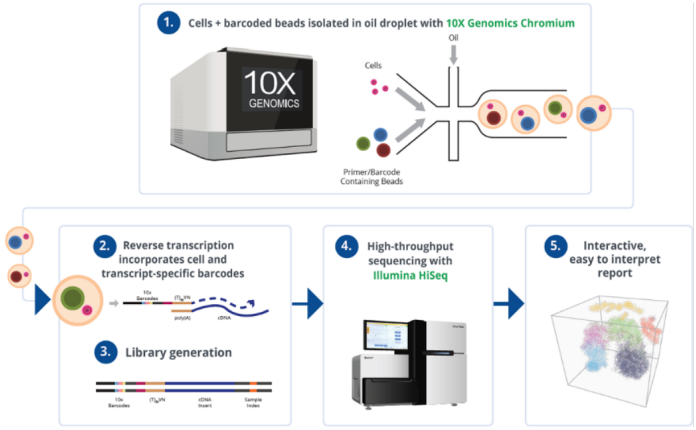
    - Epigenetic Modifications Beyond the Coding Region
        - cohesin 黏连蛋白
            - 近着丝粒区的粘连蛋白含量较高
        - TADs: topologically associating domains 拓扑关联域
        - WAPL
        - PDS5： Precocious dissociation of sisters 5
            - associate protein of cohesin 一种粘连蛋白
        - CTCF
            - 可以划分active和inactive的区域
            - 甲基化会影响CTCF的binding过程
- DNA Methylation DNA甲基化
    - 5-甲基胞嘧啶是哺乳动物中唯一发现的共价DNA甲基化
    - 影响
        - 有助于基因组的稳定性和防止转座因子
        - 基因表达
        - 基因组印记
        - x染色体失活
- 应用 application
    - De novo genovme assembly
    - Variant calling
    - Species diversity
    - ……
- Single Cell Genomics
    - Bulk RNA-seq vs Single cell RNA-seq
        - Bulk RNA-seq
            - Average transcriptional state 获得平均转录状态数据
            - Higher quality 数据质量更高
            - Cheaper
            - Less batch effect 批次效应小

        - Less technical variation 机器导致的variation影响小
        - Less bias 更小的bias
        - Standardised analysis 有标准化的分析流程
        - Best for pure cell populations 适用于细胞种类单一的细胞群
    - Single cell RNA-seq
- 标记
    - UMI 标记PCR扩增前的每一个reads，防止扩增带来的bias
    - Barcode标记细胞，相同细胞内的reads有相同的barcode
- Smart-seq2
    - 流程



        - lysis 裂解细胞
        - Reverse transcript 使用**Oligo(dT) primer**对带有polyA尾的RNA(*主要是mRNA*)进行反转录。由于使用鼠源反转录酶，末尾会加上Cs
        - Template switching 模板置换，使用**TSO (template-switching oligo) 引物**合成了cDNA的二链，从而置换了与一链cDNA互补的RNA。
        - Preamplification 预扩增，只进行几轮的PCR扩增
        - DNA fragmentation and adaptor ligation 使用高活性Tn5转座酶break DNA并添加adapter
        - Gap repair 修复gap
        - DNA purification DNA 纯化
- 10x Genomics

- 流程



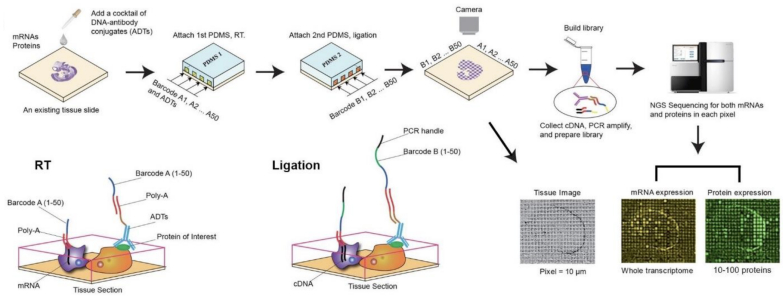
  - 使用 barcode和基胞分流仪分离单细胞
  - 逆转录
  - 建立Library
  - Hiseq高通量测序

- Smart-seq2 vs 10x Genomics

Smart-seq2	10x Genomics
Low throughput: 10^2-10^3 cells	High throughput: over 10^2-10^6 cells
Labor intensive	Labor free
Higher quality	Lower quality
Not strand specific	Strand specific
Highly abundant-transcripts bias	Cell capture rate: 65%

- Dimensionality Reduction 降维
  - 由于RNAseq的发展，对于每个细胞的基因表达谱的描述使得我们最终的数据集将有数百到数千个维度。由于细胞是分散且独立的，高维图无法告诉我们任何信息。目前主流的降维方式是t-SNE和UMAP，如果对参数有深度理解， 两者的降维结果是相似的
  - 步骤
    - 计算高维空间中一个细胞和一定数量的邻居之间的“距离”
    - 确保在降维后这些距离是相似的
  - t–Stochastic Neighbourhood Embedding (t-SNE)
    - 使用高斯概率函数计算细胞

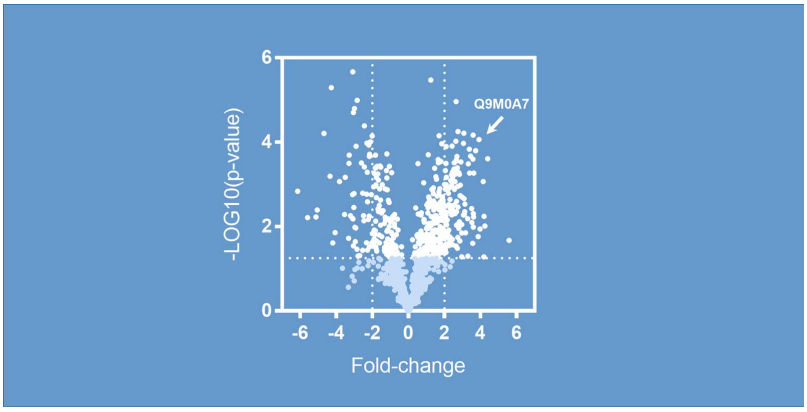
        - groups间的距离没有任何意义，groups内的距离可能有意义，但是距离的长短没有参考价值
    - Uniform Manifold Approximation and Projection (UMAP)
        - 构建拓扑结构
- Single Cell ATAC-seq
    - 只需要几百个细胞的情况下检测开放性染色体区域
    - 理想的profile
        - 发展性组织
        - 肿瘤组织
        - biopsies 活组织样本
- Spatial Multi-Omics
    - 单细胞分析技术通过将细胞脱离其生物环境而牺牲了关键信息
    - Spatial transcriptomics 空间转录组学
        - mRNA使用条形码寡核苷酸从完整的组织中捕获，该寡核苷酸可以匹配样本中的特定位置
    - DBiT-seq (Deterministic Barcoding in Tissue)
        - 使用barcode标记数千个mRNA转录本，使用oligonucleotide-tagged antibodies寡核苷酸标记的抗体标记蛋白质
        - workflow



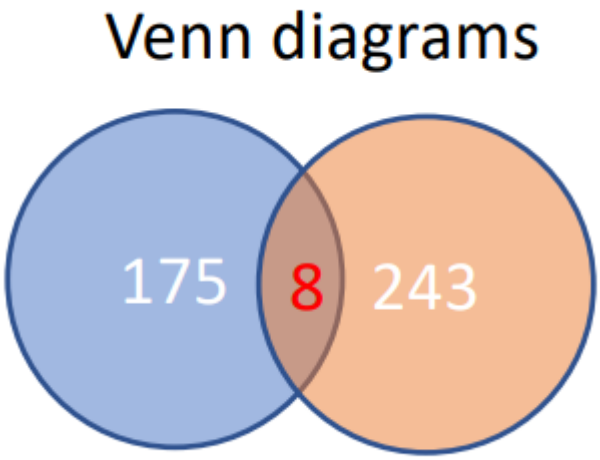
- Functional genomics
    - what is functional genomics
        - Analyses are performed on a “genome-wide” scale 是基于全基因组的分析
        - Reverse genetics -> from genomic to phenotype 从基因组出发，研究表观变化
    - application
        - identification of gene function for all genes in all organisms 识别所有物种的所有基因的功能
        - Improvement of crop yields, plant diseases, agrobiotechnology for drug production 应用于植物学，提高产量，减少疾病
        - Cancer, identification of drug targets 确定癌症、药物的靶目标
        - mechanisms of antibiotic resistance 抗生素耐药性机制的研究

- functional genomic问题的基本问答
  - How many genes are essential for the viability and optimal fitness of human cells? 人类实现生存和最佳适应性最少需要多少基因?
    - approximately 2000 genes are essential (under standard cell culture conditions and when summing up essential genes in a number of cell lines) 在标准传代条件下 大约为2000个
  - Are essential genes in human cells also essential in other organisms? 人类essential的基因是否在其它物种中是必须的?
    - many essential genes in human cells are also essential in yeast 人类essential的基因在酵母中同样是essential的
    - essential genes are more evolutionarily conserved essential的基因在进化时通常是保守的
  - What are the functions of essential genes in humans? essential的基因的功能有哪些
    - enriched in translation, transcription and DNA replication, but not signalling 转录、翻译以及DNA复制，没有信号传递功能
  - What are the differences in the set of essential genes across different types of human cells and genotypes?不同类型的人的基因类型和essential基因有什么不同
    - A core set of common essential genes (~700 genes) exist 核心的共同基因大约有700个
- Model organisms
  - Functional genomics tools
    - RNAi RNA敲除酶
    - CRISPR/Cas9
    - Gene deletion libraries
    - Over expression (cDNA/ORF)
  - why we need it?
- Chemical genomics
  - what
    - not pure functional genomics in the strictest sense 不是严格意义上的功能基因组学
    - any study directed at gaining a holistic understanding of how small molecules interact with cells 研究小分子与细胞的相互作用
    - phenotypic drug screens do not reveal drug targets,but can be combined with sgRNA/RNAi tools 无法识别药物的靶向但可以与sgRNA/RNAi结合使用
- Typical data analysis for functional genomics
  - Volcano plots 火山图
    - plotting a measure of the statistical significance of a change (e.g., p-value) on the y-axis, versus the magnitude of the change (fold-change) on the x-axis. ----y轴表示统计学意义上某个变化的重大程度，x轴表示变化的幅度

        - 可以快速选出变化幅度大，同时在统计学上也有差异的点（蛋白质，基因等），目标点应在图的右上或左上角
        - 示例



    - Gene set or GO enrichment 富集分析
    - Network analysis
        -
        - 示例
    - Venn diagrams
        - 示例

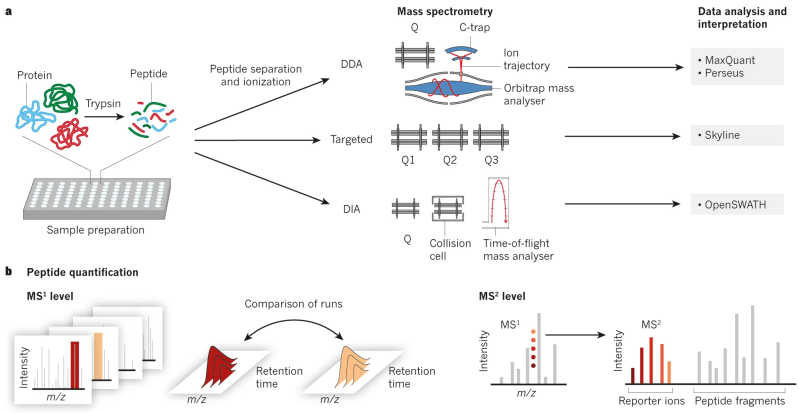


    - Distribution of samples/hits
        - 示例
    - Binding motifs
        - 示例
    - cluster by t-SINE, PCA and UMAP
        - 示例
- Shortcoming
    - Functional redundancy of genes prevents identification of observable phenotype 功能性冗余可能会掩盖可观察到的基因的表现型，如，敲除了基因x，但基因y与基因x有部分功能

重叠，导致重叠部分的的功能在对比KO和control组时无法观察到区别

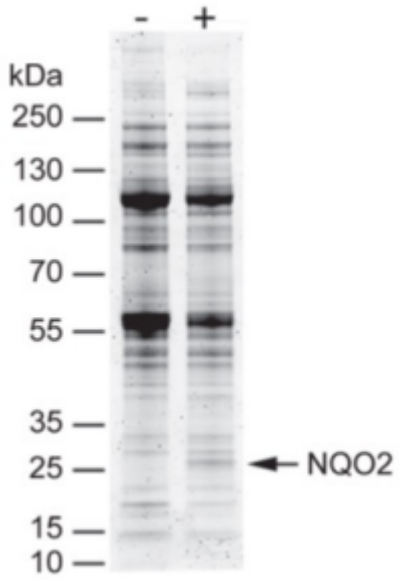
- Only a limited view on gene function as phenotype may be heavily dependent on environmental conditions 基因的表现型受到环境的影响，所发现的gene function可能是有限的
- Information overload and multiple-testing problem due to large-scale experimentation 大规模的实验可能会导致信息过载，或相同实验被多次重复
- Independent validation required? 独立的验证要求?
- Your results are only as good as your measurements! 获得的结果只在你的评价范围内是good的?

- Proteomics
  - background
    - 现状：目前蛋白质组学还没有产生临床认可的生物标志物。所有的protein-based生物标记物都是基于screening和antibody
    - 意义：蛋白质和代谢标记物更有利于研究人类的生理学
    - 困难：蛋白质不像DNA可以数字化的储存（ATGC），扩增和纯化也暂时没有标准化的方法
  - mass spectrometry
    - 质谱分析法通过mass-to-charge (m/z) 的比例，识别和定量分子
    - 质谱分析给确定蛋白质的相对量和绝对量带来了可能性。但不是每个spec experiment 都是定量的
    - common mass analyzers 常见的质谱分析仪
      - time-of-flight [TOF]
      - orbitraps
      - quadrupoles
      - ion traps
    - Tandem mass spectrometry (MS/MS) offers additional information about specific ions. 串联质谱分析可以提供特定ions的额外信息　A two step process is also called tandem mass spectrometry
    - selected ions在第一轮筛选中通过m/z选出，后续通过一系列方式片段化
    - As biological samples are complex, liquid chromatography (LC) is commonly used before MC 由于生物样品的复杂性，在质谱分析之前，通常采用液相色谱法(LC)
  - protein identification work
    - workflow
      - mass spectrometry

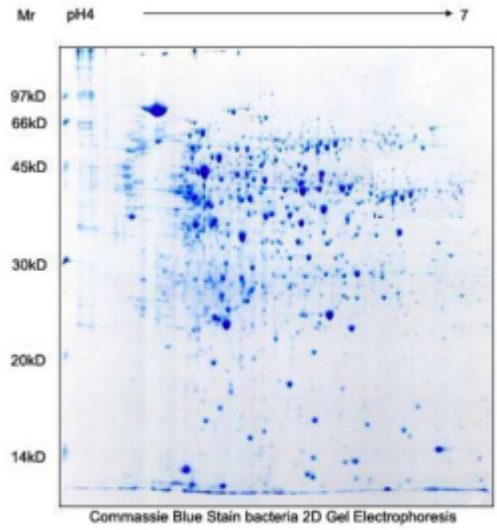


- Trypsin digestion generates smaller peptides the mass of which can be measured using MS. 使用特异性酶（如：胰蛋白酶）将蛋白质分解为更小的肽链用于质谱分析
- separate peptides 分离多肽
- These masses are compared to the calculated masses of peptides derived from in silico digested proteins. 肽链的质量与从电脑模拟消化蛋白中获得的肽的计算质量相比较，确定肽的种类
- if many peptidesmatch the expected peptides derived from in silico digest, this is taken as positive identification of a protein. 若多个肽链与某个电脑模拟的消化肽链吻合则可以判定有一个蛋白质的阳性结果

- 潜在的问题
  - 在某个细胞或者复杂组织中，可能只有某个低表达量的蛋白质被识别出。故蛋白质的识别是一个具有一定mass/charge ratio的多肽的概率的过程
  - 虽然mass measurement的误差很小，过小的mass tolerance会导致丢失目标蛋白
- 是否需要separate蛋白质
  - good for identification of individual proteins
    - 1D gel

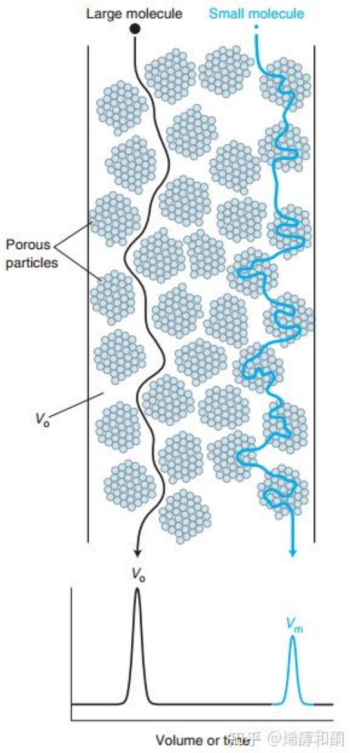


    - Two-Dimensional Gel Electrophoresis （根据性质，进一步分离多态，如PH）



Commassie Blue Stain bacteria 2D Gel Electrophoresis

- provide information for protein complexes
  - Size exclusion chromatography <https://zhuanlan.zhihu.com/p/93483415>

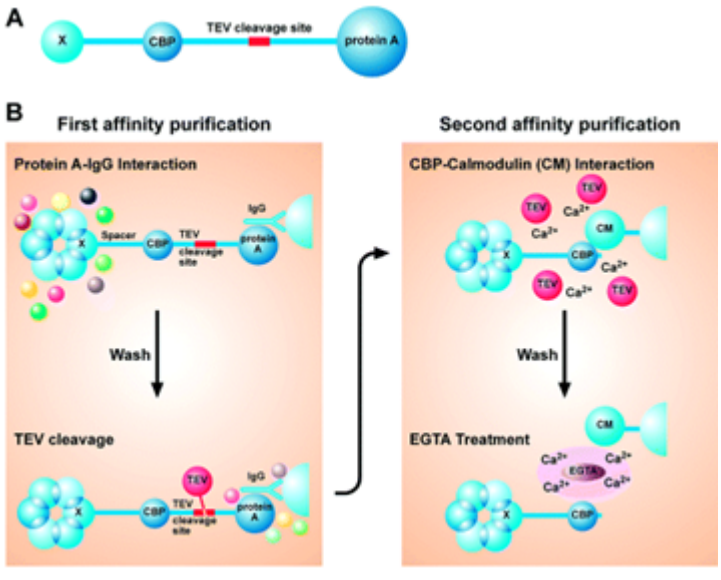


  - good for quantitative
    - Direct lysate digestionto peptides

- information we get
  - When and where proteins are expressed
  - Rates of protein production, degradation, and steady-state abundance 蛋白质的生成、降解的速率，稳定的蛋白质丰度
  - Post-translational modifications (PTMs) 翻译后的修饰

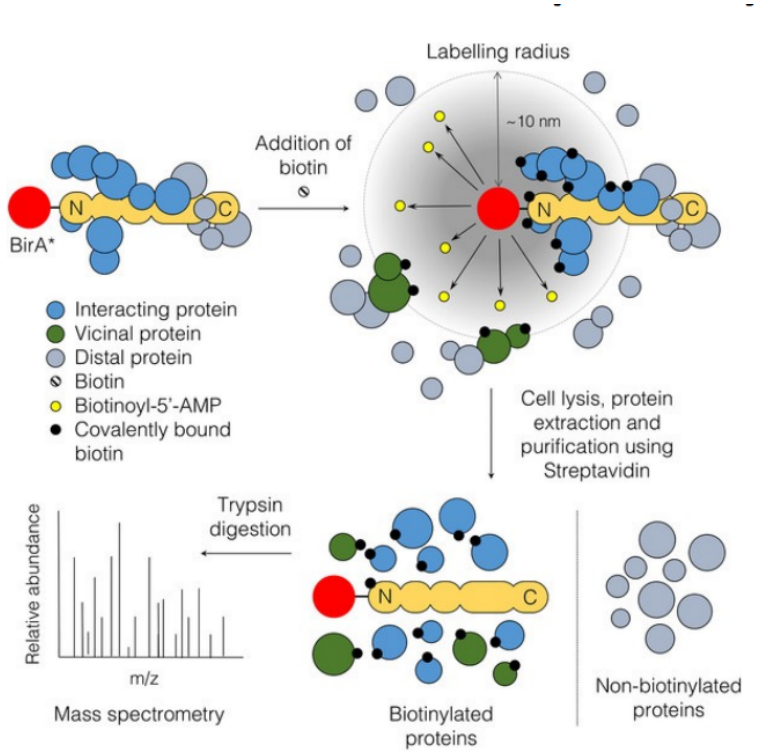
    - Localization of proteins in different subcellular compartments, or even extracellularly 定位在不同的亚细胞间室，甚至细胞外的蛋白质
    - Protein-protein interactions 蛋白质间的相互作用
- Multiplexing mass spectrometry assays
    - metabolic labeling
        - 使用Stable isotope labeling with amino acids in cell culture (SILAC)，利用内源性合成或modification的机制标记蛋白质
        - limitation
            - expensive
            - efficiency lower than 100%
    - chemical labeling
- challenge
    - Deep protein coverage is not particularly quantitative due to sampling of low abundance peptides 深度
    - Almost all proteomic analyses ignore variants 无法处理variants，具有序列变异的低丰度蛋白可能会在蛋白质组学中缺失，或者由于检测到的多肽数量较少，导致蛋白质定量的准确性降低
    - It is also possible that a mutation is identified as a post-translation modification(PTM) due to the same difference in mass 如果一个突变导致了其在mass上的变化，他可能会被识别为一种翻译后修饰
    - lack of effective algorithms and cloud computing to find known protein variants 缺少高效的寻找已知蛋白质variant的算法和云计算
    - Huge dynamic range 不同蛋白质在细胞内的含量可能差异巨大
    - MS of membrane proteins is difficult in solubilizing, separating, and digesting due to they are hydrophobic thus lack of charged lysine (K) and arginine (R) residues 由于膜蛋白的疏水性，缺少带电荷的赖氨酸和精氨酸残疾，其在MS时会较难溶解、分离和消化
- data analysis
    - collect spectra, identifies and quantifies tens of thousands of peptides. Equals to read mapping in NGS.
        - 常用的软件 MaxQuant, Mascot, SEQUEST, ProteinProspector
- variance in data 数据间的variance
    - 实验造成的variance
        - Mass spectrometry is basically random sampling of the peptides 质谱法使用随机抽样的peptides，可以 用multiplexing减少variance
        - 实验中造成variability data的分布
            - Tissue dissection and homogenization (protein extraction) 72% 提取蛋白质的组织解剖和均化作用

        - Trypsin digestion and clean-up 3% 胰蛋白酶消化和清除
        - Run-to-run instrumental variance 16% 机器的运行差异
        - Long term (>2 week) drifts in instrument stability 8% 长期的仪器稳定性的drift
- PTM
    - 常见的PTM
    - 由于modification比较罕见，为了研究PTM，需要有enrichment富集的步骤
        - 如：研究磷酸化，通常采用固定化金属离子色谱(IMAC)或二氧化钛(TiO2)为基底，这两也可用于其它大部分的PTMs
    - data get from PTM experiment 以磷酸化为例
        - MaxQuant会输出两个文件，告知检测到的磷酸化位点
            - proteinGroups.txt
            - Phospho (STY)Sites.txt
        - 所测得的磷酸化位点，是一个概率性的识别 —> 是磷酸化在某个氨基酸上可能发生的motif p-value，很少会出现概率为1的位点
        - 注意
            - 由于蛋白质的水平可能不同，最好将PTM的水平normalized to 蛋白质水平
    - 由于磷酸酶和激酶可以在几秒或几分钟内发挥作用，PTM的水平是高度动态的
        - 常用抑制剂来保存modification
        - 意味着PTM数据可能比蛋白质水平数据的噪声要大得多
- AP-MS
    - 利用protein-protein interaction确定蛋白质的亲和性关系，默认假设蛋白质复合物是相对稳定的
    - Tandem affinity purification 串联亲和纯化
        - 概念图 <https://febs.onlinelibrary.wiley.com/doi/full/10.1046/j.1432-1033.2003.03428.x?sid=nlm%3Apubmed>



        - 步骤

        - 将Immunoglobulin G （IgG）连接至TAP tag的尾部，通过离心，分离连接了tag的蛋白
        - 使用TEV酶破坏标签，释放蛋白质
        - 将calmodulin（CaM）连接至蛋白质上的剩余部分，再次离心，分离含有CaM的蛋白质
        - 使用EGTA破环标签，释放蛋白质
- BioID
    - 概念图



    - 步骤
        - 将interest protein连接上biotin ligase，它会无差别的标记10nm范围内的其它蛋白质
        - 细胞裂解，提取，使用Streptavidin纯化蛋白质
        - 放入MS
    - 缺点
        - 由于会标记10nm范围的所有东西，一些非physically interacting的物质可能会也会被标记
- Identifying true protein-protein interactions
    - SAINT (Significance Analysis of INTeractome)
        - 使用基于无标签量化label-free的概率模型来推导出true interaction的概率
    - CRAPome
        - 用不同组的negative controls来收集最常受污染质谱数据集的蛋白质的信息
        - keratins角蛋白是最容易受污染的

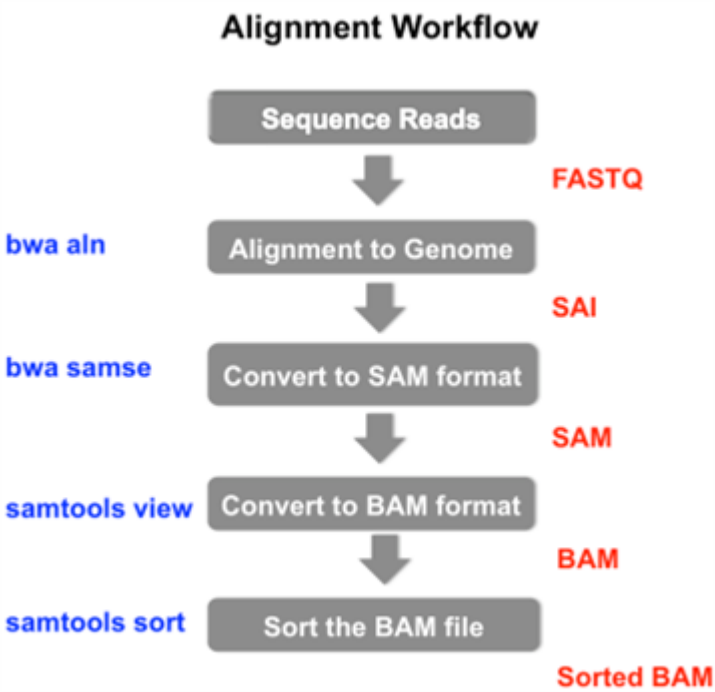
        - 通过定量，可以估计interaction的化学计量数（the molar ratio of prey proteinsand the bait protein）以及蛋白质的相对丰度
- genomics data processing
    - General steps 大致步骤
        - Data collection 收集数据
            - 确定hypothesis driven or discovery/curiosity driven
            - 确定需要多少数量的samples和genes
                - samples
                    - 是否便于统计学意义上的重复实验
                    - samples的获取是否容易
                    - 是否适合项目的预算
                - genes
                    - Real time PCR 适用于基因数量少的实验
                    - Microarrays 适用于基因数量多的实验，但sensitivity较PCR而言较低
                    - RNAseq 适用于深度分析的实验以及基于发现的项目，如：识别新的转录本，研究非编码RNA
            - 确保样本的随机性，避免batch effect批次效应
            - 从公共数据库获取数据

Consortium	what is it for?
ENCODE	Transcription factor binding sites, gene expression and epigenomics data for cell lines
Epigenomics Roadmap	Epigenomics data for multiple cell types
The cancer genome atlas (TCGA)	Expression, mutation and epigenomics data for multiple cancer types
1000 genomes project	Human genetic variation data obtained by sequencing 1000s of individuals

        - Data quality check and cleaning 检查数据质量
            - FASTQC
            - Data cleaning
                - trimmomatic
                - cutadapter
        - Data processing 数据处理
            - reference genome
                - genome version 基因版本
                    - hg19
                    - hg38
                    - GRCh38
                - genome annotation

- *UCSC refSeq*
- *Ensembl*

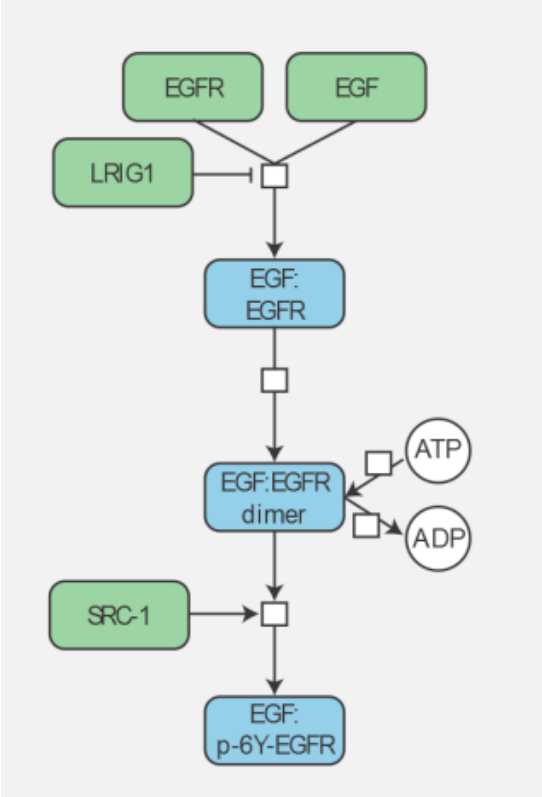
- alignment
  - Bowtie vs BWA vs STAR
  - parameter
  - workflow



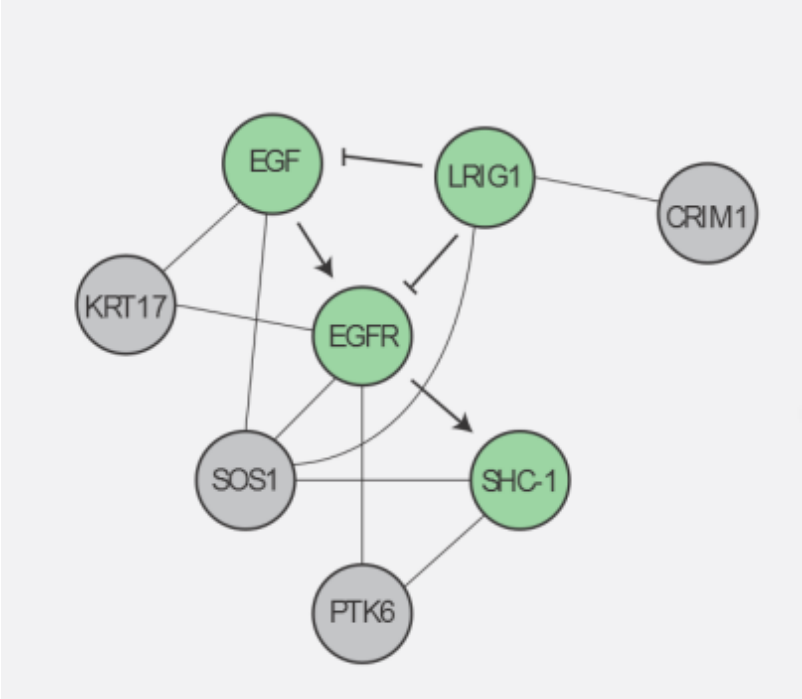
- PCR duplicate
  - UMI可以防止PCR的重复更加具有识别性
  - 以往的NGS的库中没有加入UMI
  - normalization
    - RPKM/FPKM vs TPM
      - RPKM (Reads Per Kilobase Million)
        - The sequencing depth is normalized (the 'million' part) 测序数量 normalized to 一百万
        - The length of the gene is also normalized (the 'kilobase' part) 所有基因的长度都被normalized to 一千个碱基对
        - method
          - raw data / sequencing depth -> / read length
      - FPKM (Fragment Per Kilobase Million)
        - using to deal with paired-sequencing 防止paired-sequencing数据中的片段被重复计算
      - TPM (Transcript Per Million)

- 可以确保每个重复中的gene的总数相同，适合用于统计某个基因reads的占比
- Exploratory data analysis and modeling 分析数据
  - normalization vs standardization
    - normalization
      - x轴和y轴最大值均为一，将整体数据放入这个1*1的面积中
    - standardization
      - 面积包含正负，整体数据符合标准化分布
  - principal component analysis
    - 在多维数据中发现variance较高的几个条件
  - PCA vs tSNE vs UMAP
  - define hypothesis
  - define significance level
    - fold change
    - t-test
    - Wilcoxon Rank-Sum test
- Visualization and reporting 可视化结果并报告
  - IGV
  - UCSC genome browser
  - Reproducibility and transparency 重复性和透明性
    - 重复性
      - 操作可重复——使用given data，按照步骤可以复现
      - 生物学意义可重复——采购可比样本和遵循类似的湿/干实验室协议来复现研究结论
    - How to keep 重复性和透明性
      - 确保实验正确的控制变量并且有一定重复
      - 使用正确的统计学测试方法并在必须时听取一些意见
      - 尽可能保存原始数据的副本
      - 详细的记录过程以及参数
      - 了解并遵循描述real-time PCR (MIQE)、microarray(MIAME)和RNA-seq(MINSEQE)发布所需的最基本的信息的guidance
      - 将注释好的实验提交到功能基因组学实验的公共数据库，如ArrayExpress或GEO
- Biological analysis and interpretation of omics data
  - Pathways vs. Networks
    - workflow

    - collect genomic data 收集基因组数据，如：mRNA expression
    - normalization and score 标准化数据并打分，如：differential expression
    - Generate gene list 获得基因列表
    - learn underlying cellular mechanism by pathway or network analysis 了解潜在的细胞机理
        - visualization and identify interested gene 可视化并发现目标基因
        - understand molecular mechanism 了解分子机制
        - publish model to explain data 构建模型，解释数据
- pathways



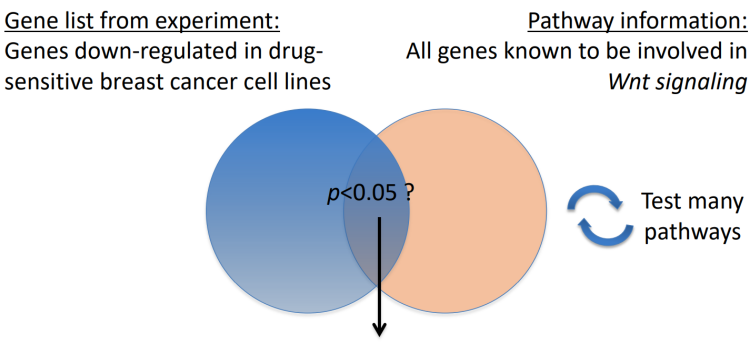
    - Detailed, high-confidence consensus 详尽，可信度较高
    - Biochemical reactions 可以展示生物化学反应
    - Small-scale, fewer genes 小范围的
    - Concentrated from decades of literature 整合于几十年的文献
- networks



    - Simplified cellular logic, noisy 简化了细胞间的作用逻辑
    - Abstractions: directed, undirected
    - Large-scale, genome-wide 范围较大
    - Constructed from omics data integration 基于组学数据的解释
- how to get the gene list
    - Molecular profiling 分子谱
        - identification -> gene
        - quantification -> gene + value
        - Ranking, clustering -> gene
    - Interactions 反应
        - Protein interactions
        - microRNA targets
        - transcription factor binding sites (ChIP)
    - Genetic screen 基因筛选
        - of knock out library查看敲除文库
    - Association studies 相关研究
        - Single nucleotide polymorphisms (SNPs)
        - Copy number variants (CNVs)
- before analysis
    - quality control your data
    - normalization
    - background adjustment
    - Use statistics that will increase signal and reduce noise 增强信号强度，减弱噪音干扰

- Make sure gene IDs are compatible with software 确保gene ID对于软件而言是可兼容的
- What do you want to accomplish with your list 思考你想用你的list作什么
  - Summarize biological processes or other aspects of gene function 总结生物学过程，或者其它方面的基因功能
  - Perform differential analysis – what pathways are different between samples?
  - Find a controller for a process (TF, miRNA) 发现某个过程的控制元素
  - Find new pathways or new pathway members 发现新的通路，或通路上的元素
  - Discover new gene function 发现新的基因功能
  - Correlate with a disease or phenotype (candidate gene prioritization) 发现是否与疾病相关
  - Find a drug 寻找药物

- method
  - Pathway enrichment analysis: summarize and compare 通路富集分析：总结与比较
    - method



      - 将实验获得的gene list与已知通路中包含的基因取交集，判断通路是否是导致实验研究问题的关键通路
    - Identifiers (IDs) are ideally unique, stable names or numbers that help track database records
      - ensemble gene ID
    - Pathways and other gene function attributes
      - Available in databases
      - pathway 通路信息
        - Gene Ontology biological process, pathway databases e.g. MsigDB, Reactome 基因本体论生物学过程以及通路的数据库
      - Other annotations 其它的信息
        - Gene Ontology molecular function, cell location
        - Chromosome position染色体位置
        - Disease association是否有疾病相关
        - DNA properties DNA信息

- TF binding sites, gene structure (intron/exon), SNPs 是否有TF的结合位点，是否有SNP
- Transcript properties 转录本信息
- Splicing, 3' UTR, microRNA binding sites 剪切，microRNA
- Protein properties 蛋白质信息
- Domains, secondary and tertiary structure, PTM sites 结构域，二三级结构，后翻译修饰位点
- Interactions with other genes 是否与其它基因相互作用

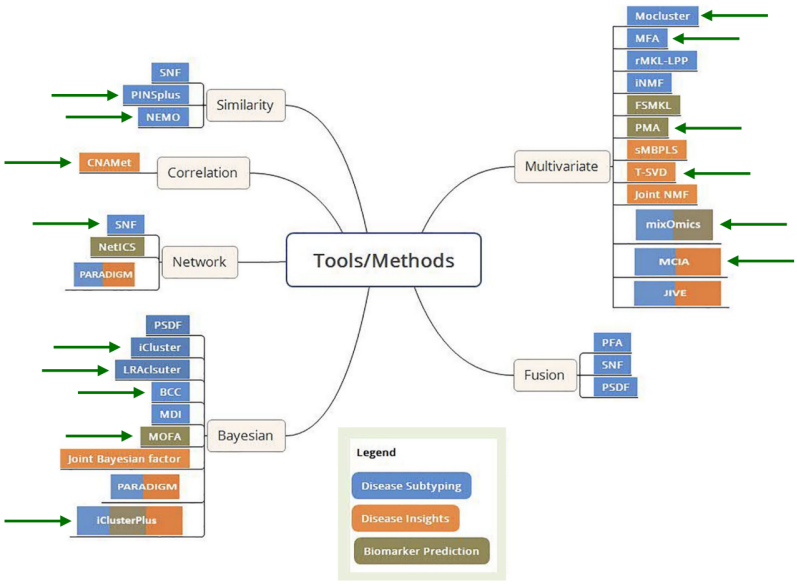
- gene ontology GO
  - what
    - ontology-- A formal system for describing knowledge 一种系统性的知识描述方式
  - cover
    - cellular component 细胞组成
      - 线粒体的结构
    - molecular function 分子功能
      - glucose-6-phosphate isomerase activity 葡萄糖六磷酸异构体的活性
    - biological process 生物学过程
      - cell division
  - Annotation Sources 注释的来源
    - manual 人为注释
      - Curated by scientists 科学家策划
        - 质量高
        - 数量少——time consuming
      - Reviewed computational analysis 通过计算分析得出的
    - electronic 未经人工验证派生的注释
      - Computational predictions 通过计算预测的
      - 质量低
      - Be aware of annotation origin 需要注意其来源
  - Enrichment analysis
    - Gene list 基因列表
      - 回答：是否某个基因在这个列表中高度富集，或被大量删除
      - Statistical test 统计学检验
        - Fisher's Exact Test (Hypergeometric test)
    - Ranked list 排序列表
      - 回答：是否有某个基因的排名出乎意料的高或低

- Statistical test 统计学检验
  - minimum hypergeometric test, GSEA
- benefit
  - easier to interpret
  - identifies possible causal mechanisms
  - predict new roles for genes 预测基因的更多功能
  - improves statistical power
  - more reproducible 可重复
  - facilitates integration of multiple data types 促进整合多种细胞类型
- Network analysis: predict gene function, find new pathway members, identify functional modules (new pathways) 网络分析：预测基因功能，识别有功能的分子
  - application
    - Gene Function Prediction 预测基因功能
    - Detection of protein complexes/other modular structures 检测蛋白质复合物或其它模块结构
    - Network evolution 升级网络信息
    - Prediction of new interactions and functional associations 预测新的相互作用或功能性信息
    - Identification of disease subnetworks 发现疾病的亚级网络
    - Subnetwork-based diagnosis 基于亚级网络的诊断
  - Network visualization and analysis 网络可视化分析
    - pathway comparison
    - literature mining
    - Gene ontology analysis
    - active modules
    - complex detection
    - network motif search
- Regulatory network analysis: find and analyze controllers

- Multi-omics and Data Integration
  - Genomics 基因组学
    - Why
      - For genotyping, to infer species evolution, to characterise genes, to find genotype-phenotype associations 研究物种进化，特征基因，寻找基因型与表现型的联系
    - How
      - Whole-genome sequencing, metagenomic sequencing, SNP typing 全基因组测序，代谢基因组学测序

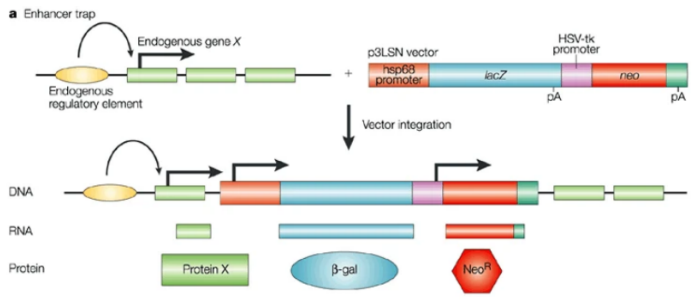
- Transcriptomics 转录组学
  - why
    - to identify differences in gene transcription between populations or due to treatments/conditions 研究基因差异性表达
  - how
    - Microarrays, RNA-seq 微矩阵， RNA测序
- Epigenomics 表观遗传组学
  - why
    - To better understand gene expression profiles 更好理解基因表达图谱
  - how
    - Bisulfite sequencing for DNA methylation 亚硫酸氢盐DNA甲基化测序，
    - ChIP-seq for histone modifications ChIPseq组蛋白修饰
    - (Small-)RNAseq for ncRNA
  - limitation
    - 转录组会由时间、环境以及细胞类型等原因变化
- Proteomics 蛋白质组学
  - Why
    - 研究蛋白质丰度及翻译后修饰
  - how
    - Mass spectrometry 质谱
    - Isotope enrichment 同位素富集
    - protein microarrays 蛋白质微矩阵
- Metabolomics 代谢组学
  - why
    - 研究某一时间点内生物样本中的某些小分子的含量
  - how
    - Mass spectrometry 质谱
  - limitations
    - 代谢物组学会随时间、环境及细胞类型变化而变化
    - 不能被电离化的分子不能被质谱仪检测出来
    - 通常无法检测到挥发性物质
- 相互作用组学
  - why
    - protein-protein interactions, miRNA-target interactions, DNA-DNA interactions, any other "set of molecular interactions in a particular cell" 蛋白质相互作用， miRNA目标相互反应， DNA相互反应， 特定细胞内的分子相互作用

    - how
        - Phage display, Yeast-two-hybrid, Affinity purification coupled to Mass spectrometry (AP-MS) 噬菌体筛选，酵母融合，AP-MS
        - DNA-protein interaction ChIP-seq
        - RNA-DNA Interactions: CLIP-seq
    - limitations
        - 不同的检测方法，亲和性
- multi-omics
    - why
        - compensate for missing or unreliable information in any single data type 弥补任何单一数据类型中缺失或不可靠的信息
        - Multiple sources of evidence pointing to the same gene or pathway are less likely to leadto false positives 指向同一基因或途径的多个证据来源不太可能导致假阳性
    - function
        - disease subtyping and classification 发现疾病的亚种及分类
        - prediction of biomarkers 预测生物标记物
        - deriving insight into disease biology 深度了解疾病的生理学
    - Methods for Data Integration
        - Tools, the one targeted by green arrow is R package



        - Correlation analysis is useful when prior knowledge of biochemical interactions is lacking 当缺少先验的生物化学反应的知识时，可以使用correlation
        - LRAcluster and IClusterPlus可以聚类整合甲基化及基因表达数据
        - mixOmics可以多元分析，数据探索，降维和可视化
- Challenges in Omics Data Integration
    - Preprocessing steps

        - lack of universal standards
        - Data filering
        - Systematic normalisation
        - Removal of batch effect
        - Quality check
    - Heterogeneity and large data size lead to computational intensity 异构性和大量的数据导致计算性能的困难
    - Benchmarking studies to evaluate the performance of different methods and tools 对标研究以评估不同方法和工具的性能
- Spatial Information by Sequencing
    - Enhancers
        - Cis-regulatory elements
            - can act with promoters over long distances 可以能与启动子在长距离下相互作用
        - enhancer trap

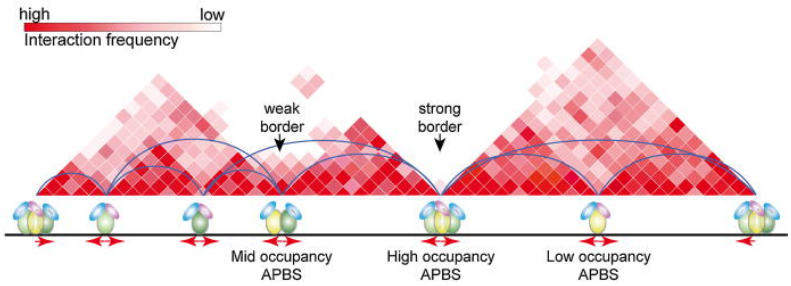


    - Chromatin Conformation Capture (3C)
        - 概述 <https://zhuanlan.zhihu.com/p/349659624>

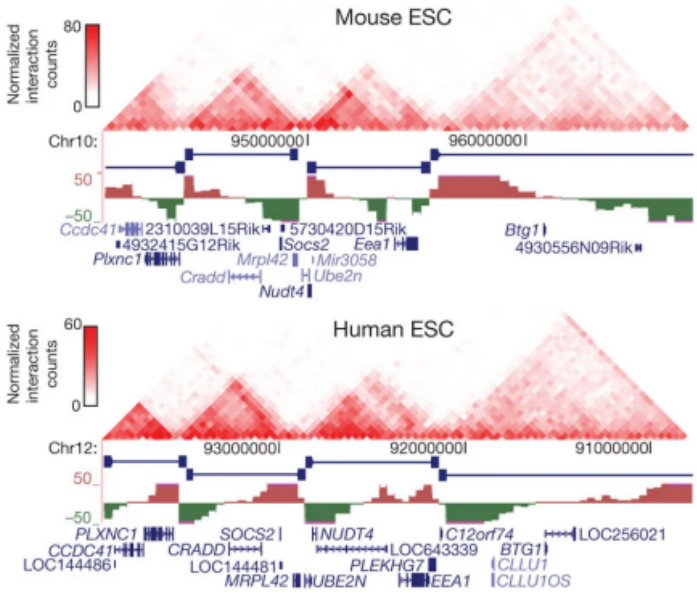
3C	4C	5C	Hi-C
One-by-one	One-by-all	Many-by-many	All-by-all
			• Biotin labelling of ends • DNA shearing 
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing

        - step 步骤
            - cross-link DNA 用交联的手段连接 Enhancer 和 promoter结合的两个DNA
            - cut with restriction enzyme 用限制性酶切割结合位点的左右两边
            - ligate DNA and digest proteins 连接两段DNA成环并消化互相连接的蛋白质
            - break the loop and get fragment 解开DNA环并碎片化
            - PCR and sequencing PCR扩增并测序片段

- Hi-C
  - 目的
    - To capture the nearby DNA fragments in the 3D genome 在三维基因组中捕获DNA片段
  - all -to-all
- Topologically Associating Domains (TADs)
  - 染色质高度相互的domain（结构域）被称为 TADs

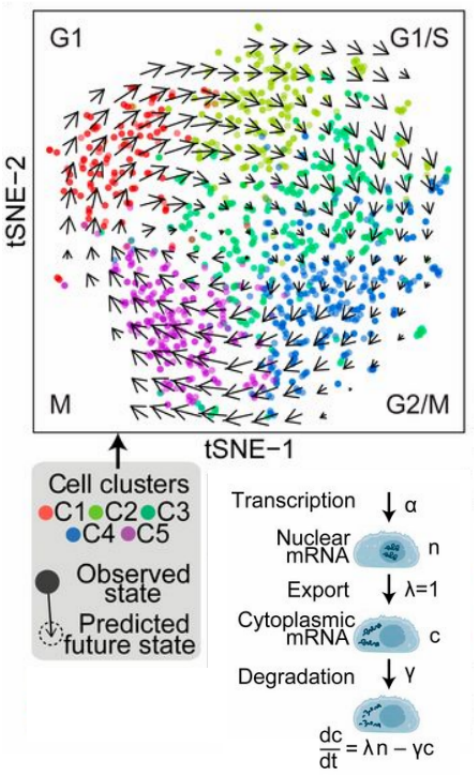


  - highly conserved across species，cell types and number of overlapping TAD boundaries 在不同物种，细胞类型中高度保守



  - boundaries are enriched in CTCF
    - CTCF is an insulator transcription factor
  - H3K27ac的活性在within TAD的活性较between TAD的活性更高
- Fluorescent In Situ Hybridisation (FISH)
  - Single Molecule FISH (smFISH)
  - Multiplexed error-robust FISH (MERFISH)
    - FISH + Super resolution microscopy + Multiplexing
      - Super resolution microscopy 可以单分子可视化
      - Multiplexing 通过使用UMI实现超过1000个分子同时操作
    - 可以用于识别细胞类型以及细胞目前所处的状态

- mRNA在细胞质中的不同阶段丰度变化示意图



- 不同mRNA在细胞中不同阶段的表达速度示意图