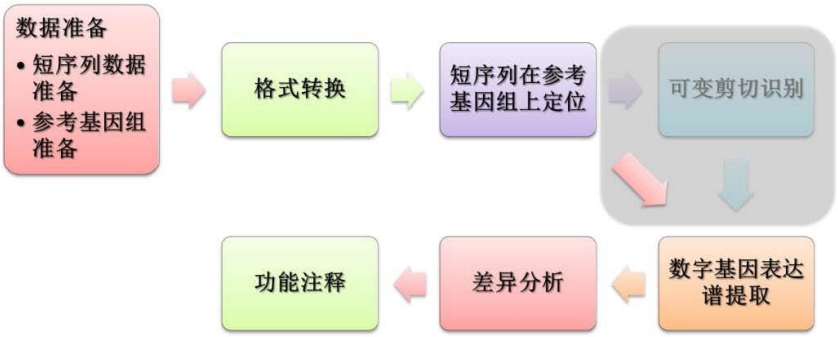
二代测序数据的获取及生物信息分析

一、分析流程图



二、主要网站及软件

NCBI 数据库

UCSC数据库

SRA 软件

bowtie软件

TopHat软件

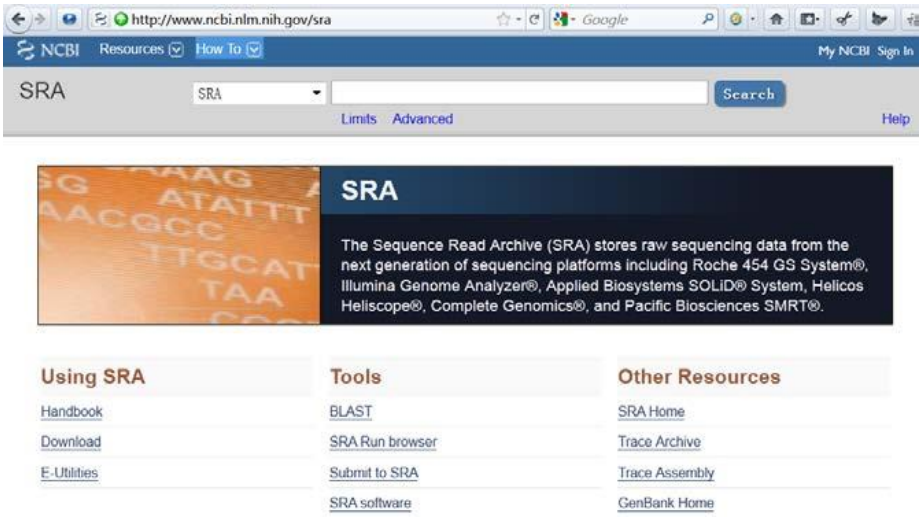
Cufflinks软件

Macs软件

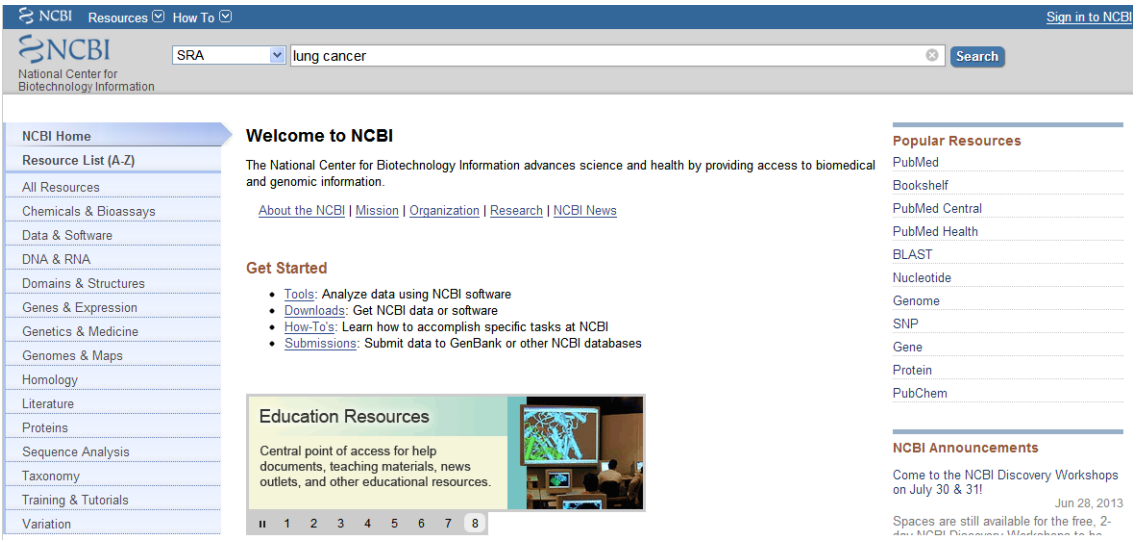
三、分析流程

1、RNA-seq数据的获取及分析

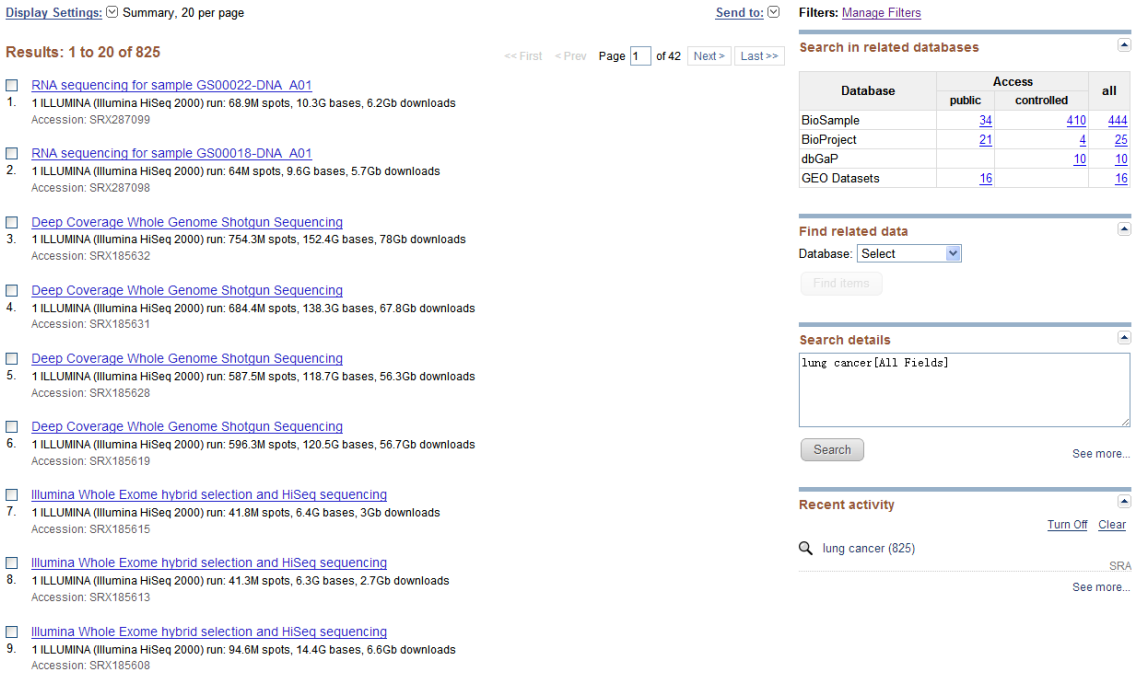
Step1：短序列数据准备：访问NCBI SRA数据库网站(http://www.ncbi.nlm.nih.gov/sra)



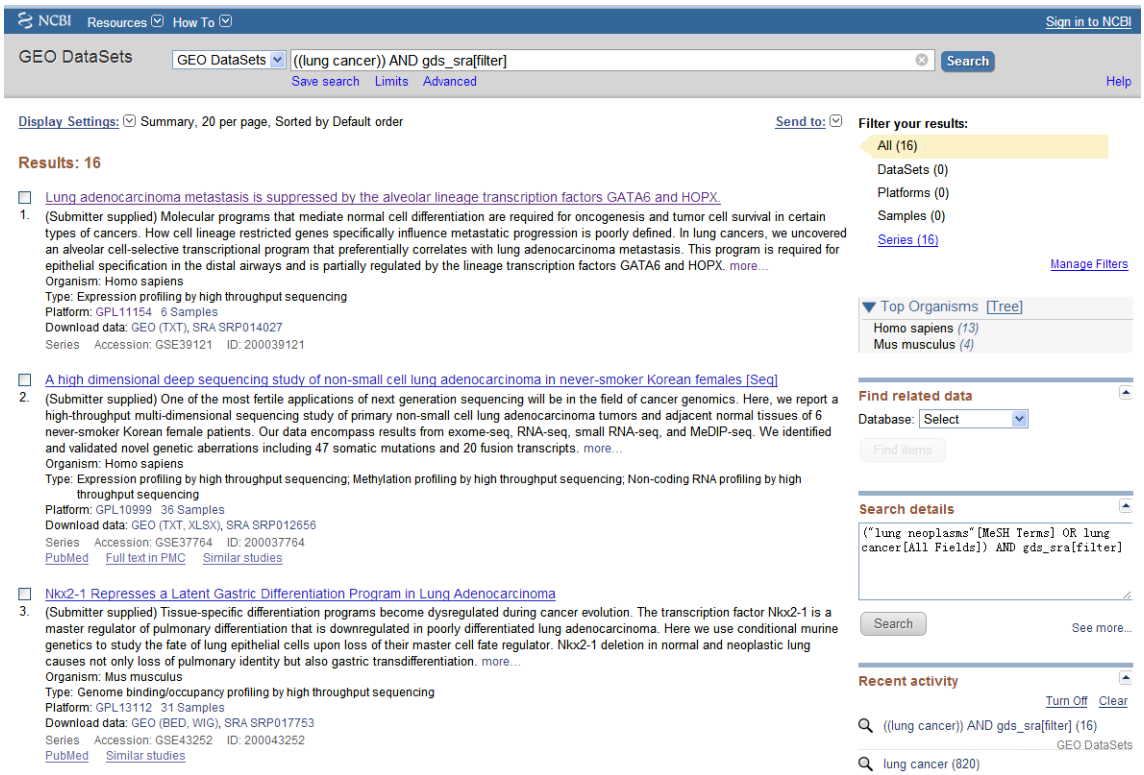
Step2：在搜索框中输入查询词，本次实验我们用“lung cancer”作为查询词。按“search”得到查询结果。



Step3：点击结果页面右侧的“Access- public-GEO Datasets(16)”，这16套数据是公共数据，可以自由下载。

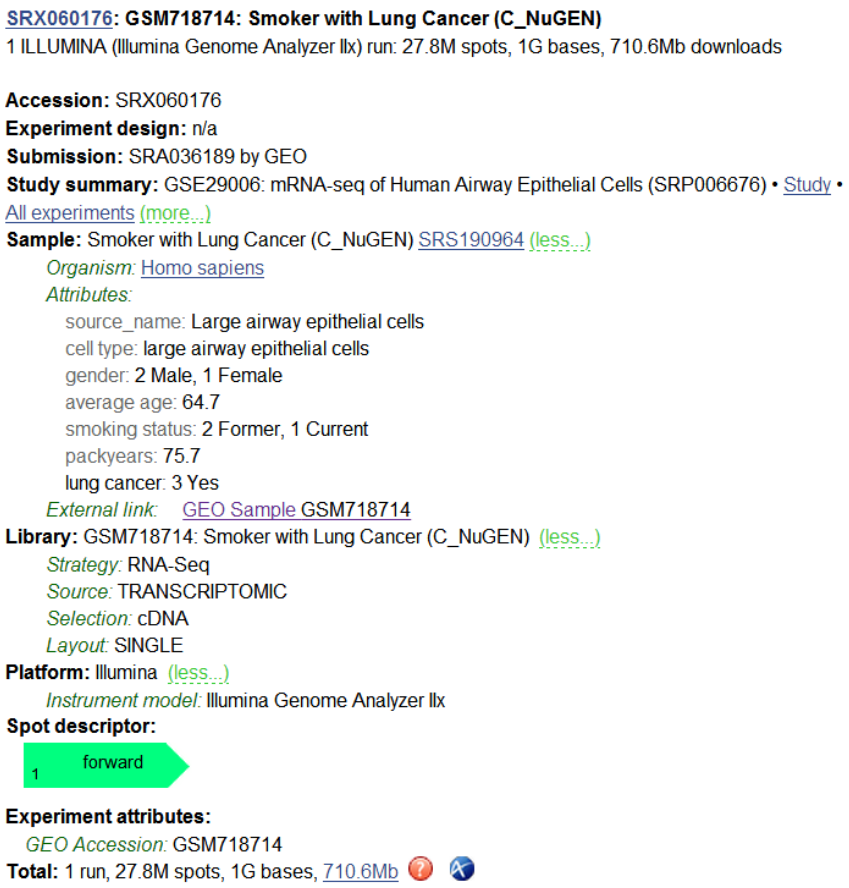


Step4：点击结果页面右侧的“Access- public-GEO Datasets(16)”，出现所有肺癌相关研究的数据列表。



Step5：点击结果页面中的第一个肺癌相关研究，出现该研究的相关描述信息。

Step6:点击本次试验页面中一个样本的信息包含关于研究（Study）、样本（Sample）、实验（Library）等信息的描述，点击查看详细信息。



SRX060176: GSM718714: Smoker with Lung Cancer (C_NuGEN)
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 27.8M spots, 1G bases, 710.6Mb downloads

Accession: SRX060176
Experiment design: n/a
Submission: SRA036189 by GEO
Study summary: GSE29006: mRNA-seq of Human Airway Epithelial Cells (SRP006676) • Study • All experiments (more...)
Sample: Smoker with Lung Cancer (C_NuGEN) SRS190964 (less...)
Organism: Homo sapiens
Attributes:
source_name: Large airway epithelial cells
cell type: large airway epithelial cells
gender: 2 Male, 1 Female
average age: 64.7
smoking status: 2 Former, 1 Current
packyears: 75.7
lung cancer: 3 Yes
External link: GEO Sample GSM718714
Library: GSM718714: Smoker with Lung Cancer (C_NuGEN) (less...)
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Platform: Illumina (less...)
Instrument model: Illumina Genome Analyzer IIx
Spot descriptor:
1 forward
Experiment attributes:
GEO Accession: GSM718714
Total: 1 run, 27.8M spots, 1G bases, 710.6Mb

Step7:确认数据可以使用以后我们可以选择下载这些数据并保存到本地。（注意事项：对于下载的每一套数据和每一个样本的数据，要做好疾病信息，样本信息的记录，以便于后续分析。）

Step8:利用“SRA software”完成短序列数据格式的转换过程，生成fastq格式的文件

Step9: 利用“bowtie”软件完成段序列在参考基因组上的定位，并且产生SAM格式的输出文件。

Step10: 利用“TopHat”软件识别段序列在参考基因组上的定位和可变剪切，并在UCSC数据库中完成可视化过程。

Step11：利用“Cufflinks”软件完成数字基因表达谱的提取过程。（输出文件为3

个transcripts.gtf、genes.fpkm_tracking和isoforms.fpkm_tracking）

2、Chip-seq 数据的获取及分析

Step1: Chip-seq 数据获取方式与 RNA-seq 数据类似

Step2：原始输入文件的格式如下：

```
ChIP.fastq
input.fastq
```

Step3: 利用“bowtie”软件完成序列的参考基因组的定位，生成 SAM 格式的输入文件，具体运行并列如下：

bowtie -p 线程数 -m 1 --sam $assembly $fq > ${fq}tmp.sam

Step4: 结果文件的数据格式转化（SAM 文件-BAM 文件-BED 文件）

1）sam 格式转为 bam 格式

samtools view -Sb ${fq}tmp.sam > ${fq}.bam

2）bam 格式转为 bed 格式

bamToBed -i ${fq}.bam > ${fq}tmp.bed

Step5：筛选特异性 map 到参考基因组单一位置的位点，具体实现命令如下：

monoExtReads.pl -w=$ext ${fq}tmp.bed > ${fq}_monoExt${ext}.bed

Step6： 输出文件的格式转化（BED 文件-BG 文件），具体实现命令如下：

bedToBedgraph ${fq}_monoExt${ext}.bed

注释：下面是生成的中间文件。

```
ChIP.fq.bam
ChIP.fq_monoExt234.bed
ChIP.fq_monoExt234.bg
hg19_Refseq.bed
input.fq.bam
input.fq_monoExt200.bed
input.fq_monoExt200.bg
```

Step7：bg 格式文件可上传至 UCSC 数据库查看，查看过程与 RNA-seq 类似。

Step8：利用“MACS”软件实现 call peaks, 具体命令如下：

macs -t Chip.bed -c input.bed - -name=yourname - -pvalue=1e-5 - -nomodel

输入文件的列表如下：

```
ChIPvsinput_negative_peaks.xls
ChIPvsinput_peaks.bed
ChIPvsinput_peaks.xls
```

peaks.xls 格式：

chr	start	end	length	summit	tags	-10*log10(pvalue)	fold_enrichment
chr1	1659994	1661447	1454	949	48	54.95	6.98
chr1	1712885	1715599	2715	1528	105	50.77	5.04
chr1	3520807	3521786	980	627	22	72.15	10.09
chr1	6211567	6212360	794	394	19	52.20	5.89
chr1	6245713	6246273	561	176	18	51.57	6.86
chr1	6673734	6674667	934	177	44	69.02	5.24
chr1	6972714	6973683	970	256	27	97.39	7.70
chr1	6987682	6988598	917	427	24	82.97	10.78
chr1	7236791	7238015	1225	588	24	64.43	11.46

peaks.bed 格式

chr1	917687	918281	MACS_peak_1	54.04
chr1	1764864	1765539	MACS_peak_2	54.48
chr1	1803786	1805898	MACS_peak_3	68.16
chr1	2101256	2102135	MACS_peak_4	52.42
chr1	2435758	2436416	MACS_peak_5	65.26
chr1	3411109	3412075	MACS_peak_6	52.10
chr1	3583070	3583869	MACS_peak_7	69.43
chr1	3822500	3823504	MACS_peak_8	53.15
chr1	6034727	6035394	MACS_peak_9	51.72

negative peaks.xls 格式

chr	start	end	length	summit	tags	-10*log10(pvalue)	fold_enrichment
chr1	1659994	1661447	1454	949	48	54.95	6.98
chr1	1712885	1715599	2715	1528	105	50.77	5.04
chr1	3520807	3521786	980	627	22	72.15	10.09
chr1	6211567	6212360	794	394	19	52.20	5.89
chr1	6245713	6246273	561	176	18	51.57	6.86
chr1	6673734	6674667	934	177	44	69.02	5.24
chr1	6972714	6973683	970	256	27	97.39	7.70
chr1	6987682	6988598	917	427	24	82.97	10.78