ISSN (Online) 2287-3406
Journal of Life Science **2021** Vol. 31. No. 4. 410~417 DOI : https://doi.org/10.5352/JLS.2021.31.4.410

ChIP-seq Library Preparation and NGS Data Analysis Using the Galaxy Platform

Yujin Kang, Jin Kang, Yea Woon Kim and AeRi Kim*

Department of Molecular Biology, College of Natural Sciences, Pusan National University, Busan 46241, Korea

Received December 11, 2020 /Revised January 22, 2021 /Accepted January 25, 2021

Next-generation sequencing (NGS) is a high-throughput technique for sequencing large numbers of DNA fragments that are prepared from a genome. This sequencing technique has been used to elucidate whole genome sequences of living organisms and to analyze complementary DNA (cDNA) or chromatin immunoprecipitated DNA (ChIPed DNA) at the genome level. After NGS, the use of proper tools is important for processing and analyzing data with reasonable parameters. However, handling large-scale sequencing data and programing for data analysis can be difficult. The Galaxy platform, a public web service system, provides many different tools for NGS data analysis, and it allows researchers to analyze their data on a web browser with no deep knowledge about bioinformatics and/or programing. In this study, we explain the procedure for preparing chromatin immunoprecipitation-sequencing (ChIP-seq) libraries and steps for analyzing ChIP-seq data using the Galaxy platform. The data analysis steps include the NGS data upload to Galaxy, quality check of the NGS data, pre-mapping processes, read mapping, the post-mapping process, peak-calling and visualization by window view, heatmaps, average profile, and correlation analysis. Analysis of our histone H3K4me1 ChIP-seq data in K562 cells shows that it correlates with public data. Thus, NGS data analysis using the Galaxy platform can provide an easy approach to bioinformatics.

Key words : Bioinformatics, ChIP-seq, galaxy, NGS

서 론

NGS (Next-generation sequencing), 즉 차세대 염기서열분석은 하나의 유전체를 무수히 많은 조각으로 분해하여 각 조각을 동시에 읽어낸 뒤, 얻은 데이터를 생물정보학적으로 처리하여 유전체 정보를 빠르게 해독하는 연구기법이다. 2000년대 중반부터 사용되기 시작한 NGS 기법은 점차 유전체 수준의 분석을 대중화시켰고, 기술 발전으로 인해 NGS 분석에 소요되는 비용은 낮아지고 있다. 그 결과 사람을 비롯한 다양한 생명체의 유전체 분석이 이루어졌으며, 사람 유전체의 다양성을 파악하기 위한 여러 연구 프로젝트들이 수행되고 있다. NGS는 대표적으로 전장유전체해독(whole genome seq), 전사체정량분석(RNA-seq), 크로마틴(chromatin)에서 DNA-단백질 결합위치분석(ChIP-seq), DNA 메틸화 분석(DNA methylation) 등에 활용되고 있다[19, 21].

초창기 NGS 기술은 염기서열을 읽는 과정에서 많은 오류를 범했고, 점차 시퀀싱 양을 늘리는 방식으로 그 문제를 해결해 왔다. 그러나 시퀀싱 양의 증가는 데이터 크기의 증가를 가져왔고, 구글 같은 기업들은 유료로 데이터를 클라우드(cloud)에 저장하는 서비스를 제공하게 되었다(https://cloud.google.com/life-sciences). 또한 NGS 데이터를 가공하고 분석하는 것도 쉬운 일은 아니다. NGS 기법이 대중화되면서 작은 연구소나 개인 연구실에서도 이 기법을 이용하여 연구를 수행하고 있지만, 컴퓨터 연산 능력이나 생물정보학에 대한 지식 부족으로 대용량의 데이터를 활용하는데 어려움을 겪고 있다. 따라서 이러한 문제를 해결할 수 있다면 일반 생물학자들도 직접 NGS 데이터를 처리하고 분석하여 생물학의 관점에서 더 많은 결과를 도출할 수 있을 것이다[7, 15].

Galaxy (http://galaxyproject.org) 플랫폼은 NGS 데이터 분석 tool을 제공하는 public 웹 서비스로, 정보학이나 프로그래밍에 대한 전문지식이 없는 연구자들에게 웹 브라우저만을 이용하여 NGS 데이터를 분석할 수 있는 환경을 제공한다. 이 플랫폼은 상당한 CPU와 디스크 공간을 제공하므로 사용자는 로그인하여 서버에 데이터 기록, 작업 순서나 흐름(workflow) 및 대규모 데이터 집합(dataset)을 저장할 수 있고[1, 2, 5, 9, 10], Galaxy의 웹 기반 그래픽 사용자 인터페이스(GUI)는 비교적 큰 데이터 분석에 필요한 모든 작업을 간단하게 수행할 수 있도록 해준다. 따라서 이러한 Galaxy GUI는 사용자가 자신의 데이터를 업로드하거나 공용 데이터베이스를 검색하는 것을 가능하게 하며, 분석 tool 선택, input 파일 선택, 매개변수(parameter) 설정 그리고 tool 실행을 가능하게 한다. 또한 Galaxy 플랫폼에서 이루어진 데이터 분석은 안정적으로 재현되며, 모든 분석 매개변수와 입력 내용은 영구히 기록된다. 그리고 사용자가 웹을 통해 분석 내용을 공유하고 게시할 수 있어서 협업적이고 투명한 분석도 가능하다.

***Corresponding author**
Tel : +82-51-510-3683, Fax : +82-51-513-9258
E-mail : kimaeri@pusan.ac.kr

ChIP (Chromatin immunoprecipitation) 분석은 크로마틴 환경에서 단백질이 유전체의 어떤 부분과 결합하는지 알아보기 위해 고안된 실험기법이다. 과거에는 PCR 또는 quantitative PCR 기법으로 항체에 의해 침전된 DNA의 특정 부분을 분석하였으나, NGS 기술의 개발로 침전된 DNA 조각 전체에 대한 분석이 가능하게 되었고, 이를 ChIP-seq (chromatin immunoprecipitation-sequencing)이라고 한다. 이는 대표적인 NGS 활용 기법으로 유전체 수준에서 전사인자(transcription factor)와 보조활성자(coactivator) 같은 단백질의 결합부위를 탐지하고, 아세틸화(acetylation)나 메틸화(methylation) 같은 히스톤 단백질의 변형(histone modification) 분포를 밝히는데 활용될 수 있다. 본 논문에서는 ChIPed DNA를 이용한 NGS용 라이브러리 제작 과정과 Galaxy 플랫폼을 이용한 시퀀싱(sequencing) 데이터 분석 과정을 단계별로 설명하고(Fig. 1), 사람 세포주 K562에서 수행한 히스톤 H3K4 monomethylation (H3K4me1)에 대한 ChIP-seq 실험의 결과를 public 데이터와 비교하여 살펴보고자 한다.

ChIP-seq 라이브러리(library) 제작 및 분석 과정

ChIP-seq 라이브러리 제작

ChIPed DNA는 점착성 말단(sticky end)을 포함하므로 end repair 기법으로 비점착성 말단(blunt end)을 만든다. 다음 과정으로 DNA 조각의 3′ 말단에 아데닌(A) 뉴클레오티드를 붙여준다. 서열이 다양한 DNA 조각들을 모두 동일한 효율로 증폭하기 위하여 양쪽 말단에 adaptor를 연결시킨다. 제대로 연결되지 않은 adaptor는 정제 과정을 통해 제거되고, 그 과정에서 DNA 절편을 원하는 길이로 선별할 수 있다. Adaptor와 상보적이면서 바코드를 포함하는 primer를 이용하여 PCR을 진행하면 ChIPed DNA의 시퀀싱을 위한 최종 라이브러리가 만들어진다(Fig. 2). 바코드는 샘플을 구분하기 위해 사용되는 짧은 서열로 한 쪽 말단에 붙일 수 있고 양쪽 말단에 서로 다른 바코드를 붙일 수도 있다. 동시에 여러 샘플을 하나의 기계에서 시퀀싱할 경우, 2개 이상의 바코드를 붙이는 것은 샘플을 분리할 때 정확성을 높여준다.

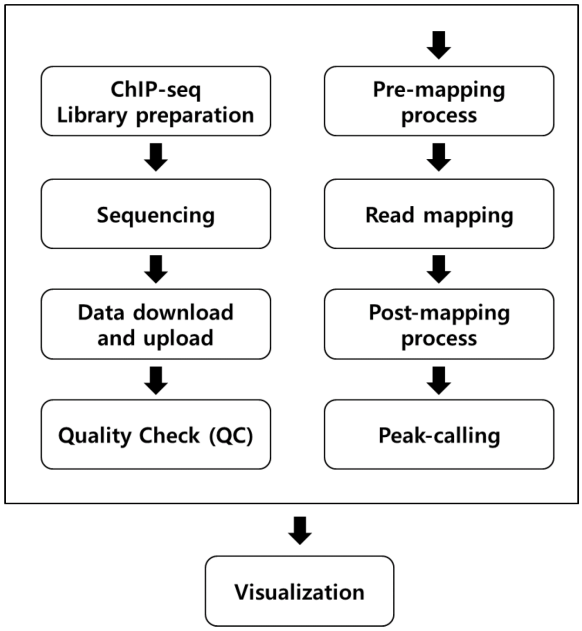

Fig. 1. Workflow of ChIP-seq analysis. The workflow represents the overall ChIP-seq analysis process. This is a standard method for analysis.

Sequencing

제작된 라이브러리는 NGS 기계를 이용하여 실제 시퀀싱을 수행하게 된다. 초창기엔 ABI, Illumina, Roche 등 여러 회사에서 NGS 기계를 생산하였으나, 현재는 Illumina 제품이 전세계적으로 사용되고 있다. Illumina에서도 NextSeq, NovaSeq, HiSeq 등 다양한 라인이 있고, 라인 별로 여러 시리즈의 기계들이 이용되고 있다. 기계에 따라 read를 읽는 길이, 방식, 개수, 용량 등이 조금씩 다르다. 시퀀싱 기계가 실제 DNA의 서열을 일부 읽어낸 결과를 read라고 하고, 하나의 DNA 조각을 single end 방식으로 한 번만 읽거나 paired end 방식으로 두 번 읽어서 read를 얻을 수 있다. 읽는 read의 길이는 일반적으

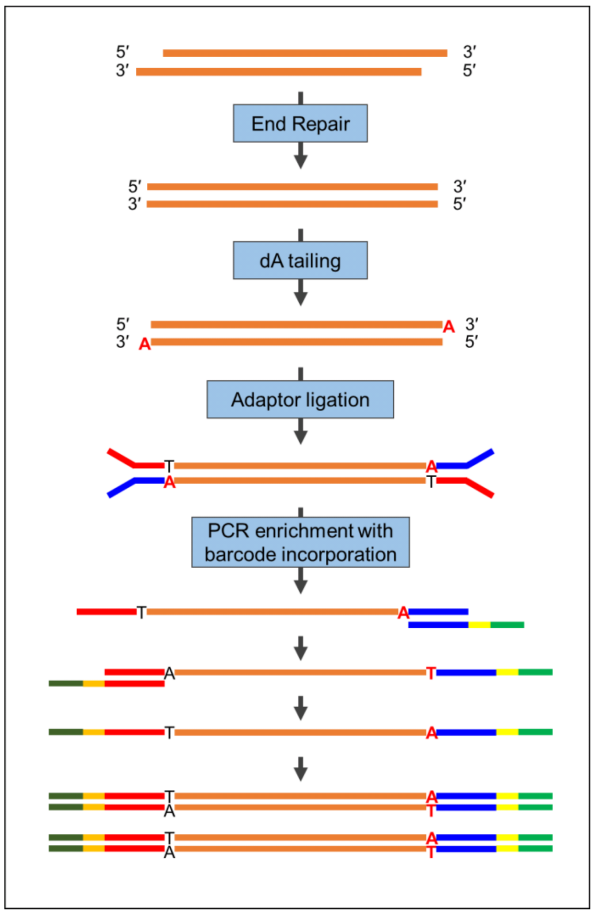

Fig. 2. Scheme of ChIP-seq library preparation. It shows the steps of ChIP-seq library preparation.

로 25-100개의 염기서열이며, ChIP-seq은 2-3천만개의 DNA 조각을 읽어서 데이터를 얻는다. 하나의 샘플에서 읽어내는 read의 길이와 개수는 실험 목적에 따라 다양하며 길게 많이 읽을수록 데이터 용량이 커지기 때문에 적절한 기준을 정하는 것이 중요하다. 본 연구에서 수행한 ChIP-seq은 3천만개의 DNA 조각을 paired end 방식으로 100개의 염기서열을 읽었으며 그 결과 6기가바이트의 데이터를 얻었다.

Galaxy composition

Galaxy 플랫폼은 여러 종류의 tool을 보여주는 영역(화면 왼쪽, Tools), 선택한 tool에 대한 분석 옵션을 보여주는 영역(화면 중앙), 현재까지 작업 내용을 보여주고 input/output 데이터의 속성을 변경하고 볼 수 있는 영역(화면 오른쪽, History)으로 나누어져 있다(Fig. 3). 기본적으로 각 tool에 대해서 input을 지정하여 해당 tool의 옵션을 설정하면 분석이 수행된다. 수행 과정은 작업(job) 대기 중, 작업 실행 중, 작업 완료, 에러 발생이 각각 회색, 분홍색, 녹색, 빨간색, 4가지 색상으로 History 영역에 표시된다.

Data download 및 upload

NGS를 수행하여 얻은 데이터 파일인 FASTQ 파일을 Galaxy 플랫폼에 업로드하거나 웹사이트 National Center for Biotechnology Information (NCBI) 또는 European Nucleotide Archive (EBI)로부터 다운로드 한 파일을 Galaxy에 업로드할 수 있다. Public 데이터는 NCBI에서 Sequence Read Archive (SRA) 파일의 SRR accession 번호를 Get Data section의 Download and Extract Reads in FASTA/Q tool에 직접 입력하면 업로드된다. 다른 방법으로 Get Data section의 EBI SRA tool을 이용하여 웹 URL입력을 통해 Galaxy 플랫폼으로 업로드할 수 있으며, EBI는 FASTQ 파일과 NCBI SRA 파일 두 가지를 제공한다.

Raw 데이터 형식

NGS를 통해 얻은 ChIP-seq 데이터는 FASTQ 형식이며, 가장 일반적인 유전자 시퀀싱 데이터 형식이다. 이는 염기서열과 함께 질평가정보인 QV (quality value)가 추가된 아스키 코드(ASCII)로 이루어진 텍스트 형태이다(Fig. 4). 보통 한 read 마다 네 개의 줄로 구성되어있다; 1) '@'로 시작하는 서열 아이디(sequence identifier) 와 시퀀싱 관련 정보 서술, 2) 해독한 염기서열 데이터, 3) '+'로 시작하며 기타 설명(생략 가능), 4) 해독한 염기서열 데이터 값의 QV를 표시하며, 염기서열 데이터와 길이가 같다. QV는 프레드 수치(Phred score)로 나타내고, 프레드 수치 Q_{Phred}는 $Q_{Phred} = -10 \log_{10}P$로 정의된다. 여기서 P는 염기를 잘못 해독할 확률로 오류율(error rate)이라 부른다. 식에서 알 수 있듯이 프레드 수치가 클수록 염기의 정확도는 높으며, 오류율은 그 반대이다. Q10은 10%의 오류율, Q20은 1%의 오류율, Q30은 0.1%의 오류율을 의미한다. 시퀀싱 장비 마다 고유한 데이터 생산 메커니즘에 따라 값이 도출되기 때문에, 서로 다른 기종의 QV를 그대로 비교하는 것은 오류의 위험성이 크다.

Quality check (QC)

QC는 raw data의 read quality를 확인하기 위한 과정이다.

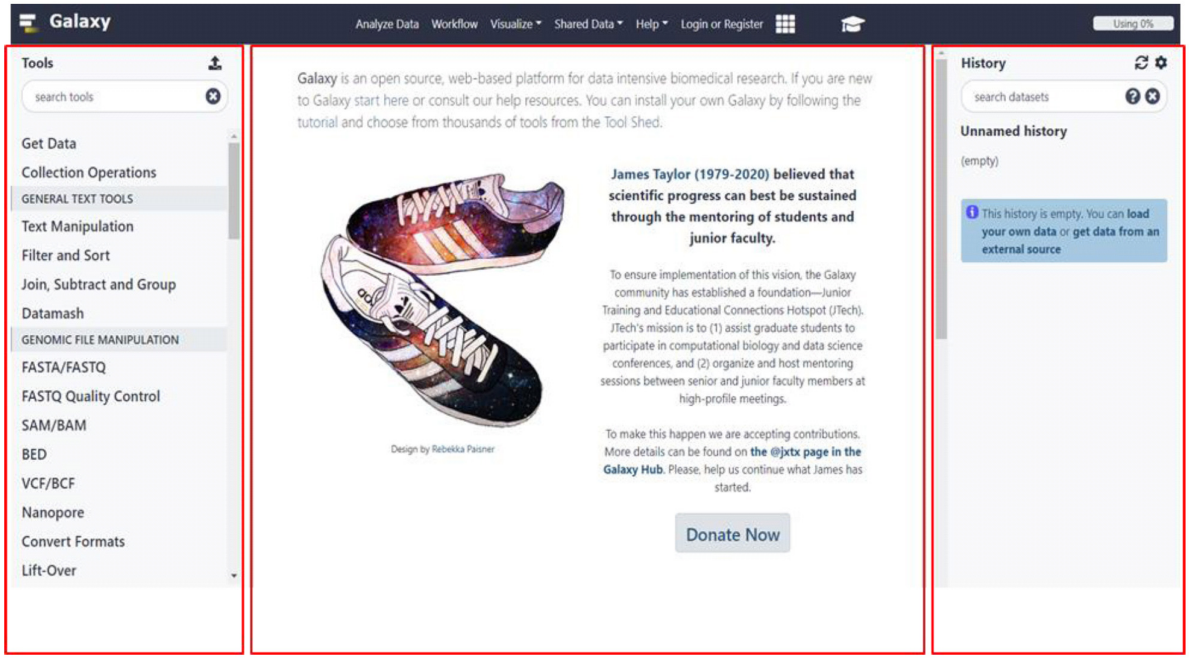

Fig. 3. Composition of Galaxy platform main. The main is composed of three regions; Tools region (left), option screen region (middle), History region (right).

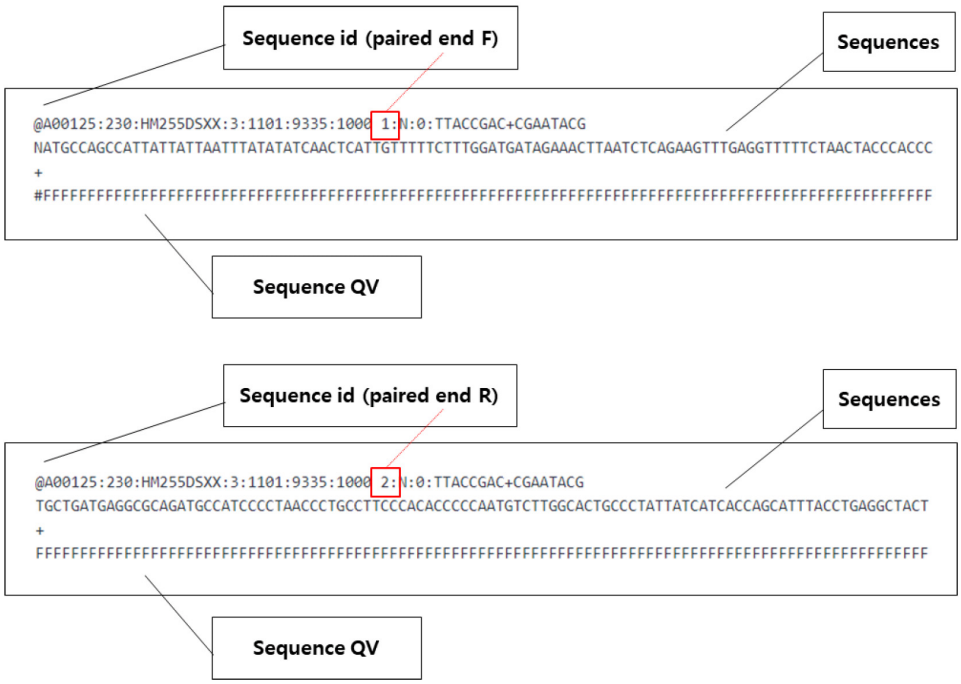

Fig. 4. Raw data format obtained from paired end sequencing. The text-based FASTQ file is a general sequencing data format. It consists of four lines per read.

일반적으로 read의 길이와 수를 검사하게 되고 오염 또는 잘못 읽은 염기서열이나 낮은 quality의 염기서열이 있는지 찾는다. QC는 FASTQ Quality Control section의 FastQC tool을 이용하여 진행한다[3]. Input 파일은 FASTQ, SAM 또는 BAM 파일 형식이어야 하고, 결과에서 11가지 항목을 확인할 수 있다. 통과(pass)는 초록색 브이 표시, 경고(warning)는 노란색 느낌표, 실패(failure)는 빨간색 엑스 표시로 나타난다. 11가지 항목은 다음과 같다. 1) Basic statistics - 분석된 파일에 대한 몇 가지 간단한 구성 통계를 보여준다. 2) per base sequence quality - read의 각 위치에서 모든 base의 quality 값의 범위를 보여준다. 3) per tile sequence quality - read의 각 위치에서 illumina flowcell의 시퀀싱 quality 정도를 보여준다. 4) per sequence quality scores - 모든 염기서열에 대한 quality 값의 분포를 나타내고 read 당 평균 quality 값을 보여준다. 5) per base sequence content - read의 각 위치에서 모든 염기의 비율을 보여준다. 6) per sequence GC content - 각 염기서열의 전체 길이에 따른 GC content를 측정하고 이를 모델링된 정규분포와 비교하여 보여준다. 7) per base N content - 제대로 읽히지 않은 염기서열의 경우 N으로 대체되고 read의 각 base 위치에서 N 비율을 보여준다. 8) sequence length distribution - read 길이의 분포를 보여준다. 9) sequence duplication levels - 모든 염기서열에 대한 중복 정도를 계산하여 비율로 보여준다. 10) overrepresented sequences - 전체의 0.1% 이상을 차지하는 염기서열을 나타낸다. 11) adaptor content - 제거되지 않고 남아있는 adaptor를 보여준다.

Pre-mapping processes

NGS를 통해 얻어진 ChIP-seq raw 데이터에서 낮은 quality 염기서열은 예상치 못한 에러를 포함할 수 있으므로, 염기서열의 quality를 향상시키기 위한 filter와 trim 과정이 필요하다. FASTA/FASTQ section의 Filter by quality tool을 이용하여 quality가 낮은 염기서열을 제거하고[11], FASTQ Quality Trimmer tool을 통해 read의 끝을 다듬는 가공 과정을 실행한다[4]. Paired end의 경우 FASTQ Quality Control section의 Trimmomatic tool을 사용하여 trim 과정을 수행한다[6]. Illumina 플랫폼의 경우에는 adaptor 제거와 같이 자체적으로 부분적이나마 서열의 정확도를 관리하며, SOLiD 플랫폼은 QV가 낮은 값도 필터링하지 않고 모두 raw 데이터로 출력하여 추후 분석과정에서 처리하게 한다.

Read mapping

Pre-mapping processing 후, Mapping section의 Bowtie2 tool을 사용하여 ChIP-seq 데이터를 사람 또는 생쥐 reference 유전체에 mapping 시킨다[13, 14]. 일부 유전체는 Full 뿐만 아니라 Canonical, Canonical Female reference 같은 옵션도 존재한다. 사람 reference 유전체에서 Full의 경우는 mapping 되지 않은 미토콘드리아, 플라스미드, 다양한 샘플에서 반복적으로 관찰되지만 정확한 출처를 모르는 DNA 서열을 모두 포함한다. Canonical은 chr1-chr22, chrX, chrY, chrM을 포함하고, Canonical Female은 chrY를 제외한 염색체를 포함하므로 적합한 reference 유전체 옵션을 선택한다. Analysis mode는 기본적으로 Default setting 옵션을 선택하여 mapping을

진행하며 본 연구에서도 default 옵션으로 수행하였다. Output은 SAM 또는 BAM 파일 형식이다. SAM 파일은 sequence alignment 데이터를 담고 있는 텍스트 파일로 각 내용들은 탭으로 분리되어 mapping 정보를 담고 있다. 텍스트 파일의 문자열 형식으로 저장하여 바로 열람이 가능하며, 이를 압축하고 binary 형식으로 변환한 것이 BAM 파일이다. 따로 output 파일 형식을 지정하지 않으면 기본적인 output은 용량이 적은 BAM 파일이므로 결과 데이터를 자세히 볼 수 없다. 하지만 bowtie2 mapping statistics to the history 옵션을 선택하면 mapping에 대한 더 상세한 정보를 얻을 수 있다. 전체 read 중 정확하게 한 번 align된 read 수, 두 번 이상 align된 read 수, align되지 않은 read 수, 그리고 전반적인 alignment rate를 확인할 수 있다.

Post-mapping processes

Mapping에서 가장 큰 문제는 read가 reference 서열의 여러 위치에 정렬되거나 잘못된 위치에 정렬되는 것이다. 따라서 mapping 후 mapping quality가 낮고 다중 mapping된 read를 우선 제거해야 한다. SAM/BAM section의 Filter SAM or BAM, output SAM or BAM tool은 mapping quality score가 낮은 read를 제거해준다[16]. 여기서 중요한 옵션은 Minimum MAPQ (mapping quality) quality score이고 multi-map의 경우 score 0 값을 가지므로 MAPQ score가 적어도 1 이상 되어야 한다. 이후 같은 section의 Samtools sort tool을 사용하여 chromosome별로 정렬[16, 17], RmDup tool로 PCR 에러로 인한 duplicate read를 제거한다[18]. 이러한 과정들은 mapping의 정확도를 높이는데 기여한다.

Quality check (optional)

유전체에 read mapping이 끝난 후 가공 처리된 aligned read의 QC를 수행한다. 이 때 사용되는 tool은 앞선 QC과정과 마찬가지로 FASTQ Quality Control section의 FastQC tool을 이용한다.

Peak-calling

ChIP-seq 실험에서 enriched binding site를 찾기 위해 peak-calling 과정이 필요하다. 이를 위해 ChIP-seq section의 MACS2 callpeak tool을 사용하며, noise 보정을 통한 specificity 향상을 위해 ChIP input과 같은 control 샘플 사용을 권장한다[8, 22]. Noise는 실험적 noise와 생물학적 noise로 나눌 수 있는데, 실험적 noise는 해당 target이 아닌 유전체 상의 다른 부분이 시퀀싱된 것이다. 이를 background noise라고 부른다. 생물학적인 noise는 유전체 상의 특정 부분, 예를 들어 반복 서열(repetitive sequence)이 있는 부분은 실험적인 이유에서가 아니라 서열의 반복 횟수 등에 따라 alignment가 많이 되기 때문에 peak처럼 보이기도 한다.

Align된 read는 peak의 양 말단이므로 실제 단백질 결합 부위인 중앙으로 이동시키는 'build the shifting model' 옵션을 사용해야 하며, 이동 거리는 임의로 설정할 수 있다. 전사인자와 보조활성자는 특정 서열을 인식하여 결합하므로 좁은 범위에서 관찰되는 반면 히스톤 변형은 상대적으로 넓은 범위에서 관찰된다. 따라서 전사인자와 보조활성자는 narrow peak 옵션을, 히스톤 변형은 broad peak 옵션을 일반적으로 사용한다. 유의미한 결합 부위 탐지를 위한 false discovery rate (FDR)은 기본적으로 narrow peak는 0.05, broad peak는 0.1 이며, 대체적으로 default FDR 값을 사용한다. Peak output은 BED 파일 형식으로 만들어지며, narrow peak는 염색체 이름, peak 위치(start, end), peak 이름, genome browser에서 보여지는 score 값, strand, fold change, $-\log_{10}$pvalue, $-\log_{10}$qvalue, peak의 정점 위치를 보여주는 BED6+4형식이고, broad peak는 peak의 정점 위치를 제외한 BED6+3 형식이다.

Visualization

ChIP-seq은 크게 전사인자 결합 부위를 찾거나 히스톤 변형 부위를 알아보기 위해 수행된다. 전사인자 결합 부위에 대한 peak 길이는 수십에서 최대 수백 서열 정도지만, 히스톤 변형 부위의 peak는 길게는 100만 서열 단위까지 관찰된다. 이렇게 분석한 데이터를 시각화하기 위해서 Bigwig, Bedgraph 또는 Wig 형식의 파일이 필요하다. Bigwig 파일은 Convert Formats section의 Wig/BedGraph-to-bigWig tool을 사용하여 Wig 또는 Bedgraph 파일을 변환시켜 얻을 수 있다. 또한 deepTools section의 bamCoverage tool을 사용하여 BAM 파일을 Bigwig 파일로 만들 수 있다[20]. 이 tool은 원하는 유전체 크기, RPKM, CPM 등 다양한 normalization 방법을 제공한다. 다음은 K562 세포주에서 수행한 히스톤 H3K4me1 (Con H3K4me1) ChIP-seq 데이터를 public (Pub H3K4me1) 데이터와 비교 분석한 결과이다.

Window로 보여주기

Mapping된 read 데이터를 시각화하는 가장 기본적인 방법은 genome browser이다. 실제 유전체 전체를 분석하기 전에 직접 눈으로 데이터를 확인할 수 있는 단계로, UCSC Genome Browser 또는 Integrated Genome Browser (IGB)에서 원하는 유전체 범위를 설정하여 signal을 쉽고 간단하게 분석할 수 있다. 또한 signal을 나타내는 track의 색깔과 높이 등을 원하는 형태로 지정할 수 있으며, track 간의 추가적인 연산을 수행함으로써 control과 treatment 사이의 차이를 시각화할 수도 있다. 이를 위해 mapping된 BAM 파일을 사용하거나 Bigwig, Bedgraph, Wig 파일을 이용한다. Fig. 5A처럼 사람 베타-글로빈 좌위(β-globin locus)의 히스톤 H3K4me1 분포를 보여줄 수 있다[12].

Heatmap and average profile 만들기

모든 유전체 영역에 대한 enrichment value를 heatmap으로 시각화하거나 평균 enrichment value를 profile로 시각화한

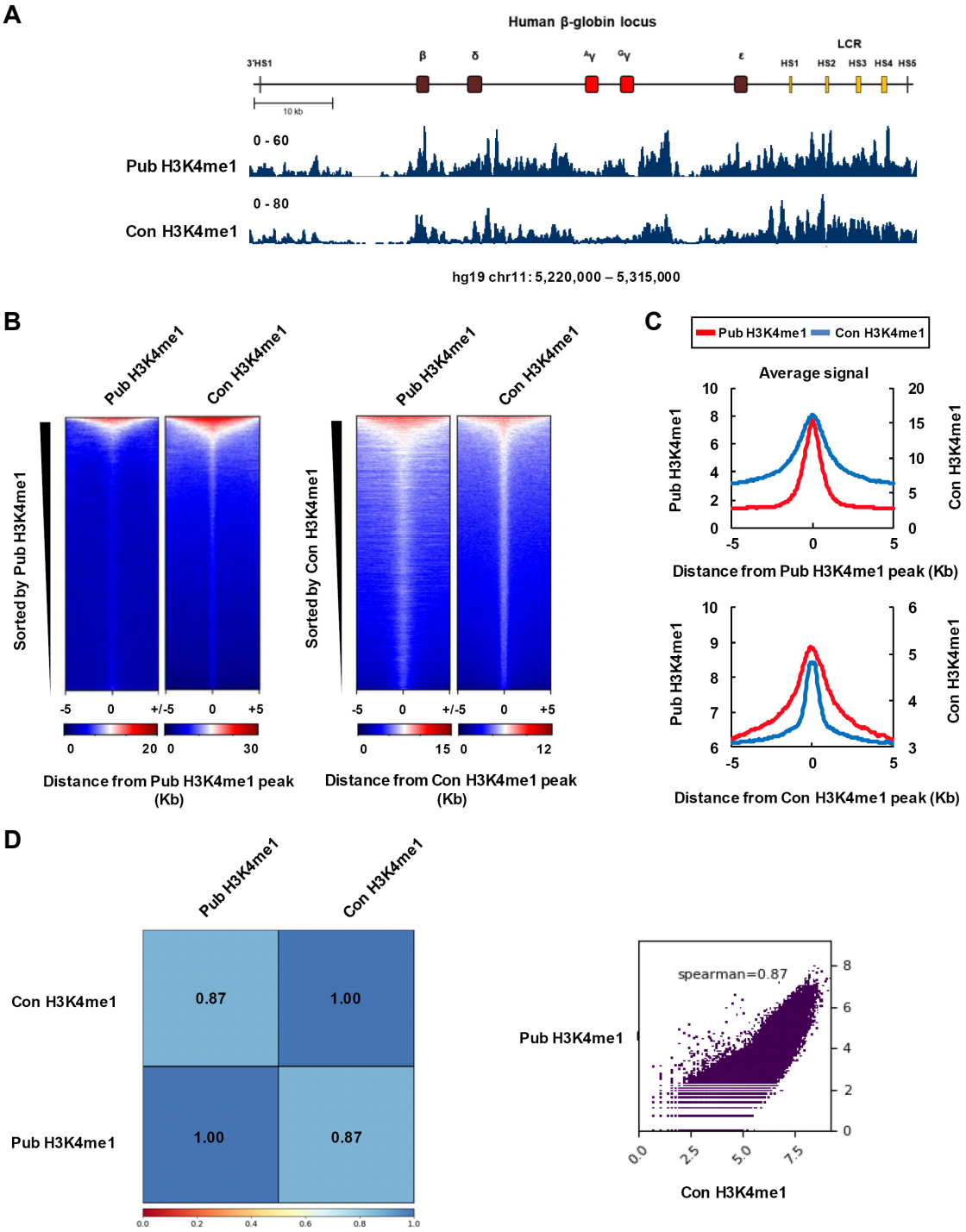

Fig. 5. Various visualization of public histone H3K4me1 ChIP-seq data in K562 cells (Pub H3K4me1) and H3K4me1 ChIP-seq data in our control K562 cells (Con H3K4me1). (A) IGB genome browser tracks show distribution of Pub H3K4me1 and Pub H3K4me1 at the human β-globin locus. (B) Heatmaps show signal enrichment of Pub H3K4me1 and Con H3K4me1 at Con H3K4me1 peaks (left panel) or Con H3K4me1 peaks (right panel). Five kilobase pairs around the center of each peak are displayed. Color scales indicate the relative signal intensity on heatmaps. (C) Pub H3K4me1 and Con H3K4me1 signals (±5 Kb from the each center) were presented by average profiles. (D) Heatmap (left panel) and scatterplot (right panel) show that Con H3K4me1 positively correlates with Pub H3K4me1. BAM files were used to generate heatmap and scatterplot using MultiBamSummary and plotCorrelation. Pairwise correlation coefficients among samples were computed by Spearman method. Pub H3K4me1 ChIP-seq data was obtained from Gene Expression Omnibus (GEO). GEO accession numbers of the data are Pub H3K4me1 (GSM788085) and Con H3K4me1 (GSE147826).

다(Fig. 5B, Fig. 5C). 이때 사용되는 tool은 deepTools section의 plotHeatmap과 plotProfile이다[20]. 두 가지 tool은 value를 계산하기 위해서 먼저 같은 section의 computeMatrix tool을 진행해야 하며, 이를 위해 genomic regions과 그 region과 관련된 score를 가지는 Bigwig 파일이 필요하다. computeMatrix는 scale-regions mode와 reference-point mode, 두 가지의 주요 mode가 존재하며, scale-regions mode는 모든 region을 같은 길이로 맞춰주고, reference-point mode는 계산된 value의 초점을 각 region의 start, end, center로 선택할 수 있다. plotHeatmap과 plotProfile tool에는 색깔, 최대/최소값 설정, 크기 조절 등 다양한 옵션이 있으며 heatmap의 경우 각 region에 대한 value 값을 토대로 region을 감소, 증가, same order로 분류할 수 있다[20].

Data correlation 분석

대부분의 실험은 반복 실험을 진행하고 반복한 실험의 유사성을 증명하거나 published data와 비교한다(Fig. 5D). 먼저 deepTools section의 multiBamSummary tool을 사용하여 두 개 또는 그 이상의 mapping된 BAM 파일에서 genomic region read를 계산한다[20]. Bins mode와 BED file mode, 두 가지 계산 모드를 선택할 수 있으며, bins mode는 동일한 사이즈의 bin에 대해 count하고, BED file mode는 특정 region에 대해 여러 bam 파일의 중복 read 수만 고려한다. 이렇게 생성된 matrix를 시각화하기 위해서 같은 section의 plotCorrelation tool을 이용한다. 이 tool은 서로 다른 샘플 간의 상관관계 값을 heatmap 또는 scatterplot으로 보여준다. 상관관계 계산은 샘플 간의 측량적 차이(metric differences)를 계산하는 pearson이나 순위(rankings)에 기초한 spearman 방법을 선택하여 사용할 수 있으며, 샘플 간의 상관관계가 강할수록 상관계수는 1에 가깝다. Con H3K4me1과 Pub H3K4me1의 상관계수는 0.87로 측정되었으며 유사한 결과라고 판단된다.

결 론

현재는 유전체학 또는 에피유전체학의 시대라 해도 과언이 아니다. NGS 기술의 발전은 생명과학뿐만 아니라 질병진단과 신약개발 등 의료 분야에도 기여하고 있으며, 기존의 연구 방법을 새로운 기법으로 탈바꿈시키기도 한다. 유전체 수준의 데이터가 증가함에 따라 대용량의 데이터를 처리할 수 있는 컴퓨팅 기술이 필요하여, 생물학, 통계학, 컴퓨터과학 등을 아우르는 생물정보학 분석기술이 요구된다. 종종 실험 기술 및 경제적인 이유로 NGS 관련 실험과 분석이 전문업체에 위탁되지만, 위탁업체의 데이터 분석은 주로 시퀀싱 결과에 대한 1차적인 분석이며, 연구자가 원하는 정보를 얻는데 한계가 있다. 또한 생물정보학적 분석 방법에 대한 이해가 부족하면 데이터 해석 과정에서 오류를 범할 수 있다. 따라서 생물학적으로 유의미한 결과를 얻기 위해 연구자 스스로 NGS 데이터를 이해하고 다룰 수 있는 능력이 필요하며, 이는 실험 디자인 및 실험 수행에도 도움을 줄 것이다.

NGS 데이터는 테라바이트급 대용량 데이터로 개인용 컴퓨터를 이용하여 처리하는데 한계가 있으며, 현재 중대형 컴퓨터와 다중 사용자 운영체제인 유닉스(UNIX) 혹은 리눅스(linux)가 많이 사용되고 있다. 그러나 이러한 컴퓨팅 환경은 프로그래밍 전문지식이 없는 연구자들이 쉽게 접근하기 어렵다. 이 문제점에 대한 해결책으로 NGS 데이터 분석을 위한 수많은 tool을 제공하는 public 웹 서비스인 Galaxy 플랫폼이 제시될 수 있으며, 이 플랫폼의 효율적 이용은 생물정보학을 전공하지 않은 연구자들도 큰 진입장벽 없이 ChIP-seq이나 RNA-seq 데이터 분석을 가능하게 해준다. 또한 변화하는 분석 기술에 따라 새로운 tool이 플랫폼에 추가되거나 최신버전으로 업데이트됨으로써 다양한 연구에 적절히 활용될 수 있다. 따라서 Galaxy 플랫폼을 이용한 NGS 데이터 분석은 많은 연구자들에게 생물정보학적 연구에 대한 손쉬운 접근 방법을 제공할 것으로 생각된다.

감사의 글

이 논문은 부산대학교 기본연구지원사업(2년)에 의하여 연구되었음.

The Conflict of Interest Statement

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A. and Goecks, J. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44**, W3-10.
2. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B. A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A. and Blankenberg, D. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537-544.
3. Andrews, S.n.d. FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
4. Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A. and Galaxy, T. 2010. Manipulation of FASTQ data with Galaxy. *Bioinformatics (Oxford, England)*

26, 1783-1785.
5. Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. 2010. Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **Chapter 19**, Unit 19.10.11-21.
6. Bolger, A. M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.
7. Celesti, A., Fazio, M., Celesti, F., Sannino, G., Campo, S. and Villari, M. 2016. *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 267-270.
8. Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728-1740.
9. Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W. J. and Nekrutenko, A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451-1455.
10. Goecks, J., Nekrutenko, A. and Taylor, J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86.
11. Gordon, A. 2010. FASTQ/A short-reads pre-processing tools. Retrieved from http://hannonlab.cshl.edu/fastx_toolkit/.
12. Kim, Y. W., Kang, Y., Kang, J. and Kim, A. 2020. GATA-1-dependent histone H3K27 acetylation mediates erythroid cell-specific chromatin interaction between CTCF sites. *FASEB J.* **34**, 14736-14749.
13. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25.
14. Langmead, B. and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359.
15. Langmead, B. and Nellore, A. 2018. Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* **19**, 208-219.
16. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.
17. Li, H. 2011. Improving SNP discovery by base alignment quality. *Bioinformatics* **27**, 1157-1158.
18. Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993.
19. Pareek, C. S., Smoczynski, R. and Tretyn, A. 2011. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413-435.
20. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. and Manke, T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165.
21. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. and Thermes, C. 2014. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418-426.
22. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. and Liu, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

초록 : ChIP-seq 라이브러리 제작 및 Galaxy 플랫폼을 이용한 NGS 데이터 분석

강유진 · 강진 · 김예운 · 김애리*
(부산대학교 자연과학대학 분자생물학과)

NGS (Next-generation sequencing), 즉 차세대염기서열분석은 유전체 수준의 방대한 DNA를 작은 절편으로 만들어서 그 절편들의 염기서열들을 동시에 읽어내는 기법이다. 현재 다양한 생명체의 유전체 염기서열 분석부터 cDNA (complementary DNA)나 ChIPed DNA (chromatin immunoprecipitated DNA)를 분석하는데 이 NGS 기법을 사용하고 있으며, 이 때 얻어진 데이터를 적절히 처리하고 분석하는 일은 생물학적으로 유의미한 결과를 얻기 위하여 중요하다. 하지만 대용량 데이터의 저장 및 활용, 그리고 컴퓨터 프로그래밍 바탕의 데이터 분석은 실험을 수행하는 일반 생물학자들에게 어려운 일이다. Galaxy 플랫폼은 다양한 NGS 데이터 분석 tool을 무료로 제공하는 웹 서비스이며, 생물정보학이나 프로그래밍에 대한 전문지식이 없는 연구자들에게 웹 브라우저만을 이용하여 데이터를 분석할 수 있는 환경을 제공한다. 본 논문에서는 ChIP-seq (chromatin immunoprecipitation-sequencing) 수행을 위한 라이브러리 제작 과정 및 Galaxy 플랫폼을 이용한 ChIP-seq 데이터 분석 과정을 설명하고, K562 세포주에서 수행한 히스톤 H3K4me1 ChIP-seq 결과가 public 데이터와 일치함을 보여준다. 따라서 Galaxy 플랫폼을 활용한 NGS 데이터 분석은 생물정보학에 대한 손쉬운 접근 방법을 제공할 것으로 기대된다.