VELVET assembler를 이용한 Illumina pair-end seuence들의 assembly

Illumina사의 Solexa system을 이용한 *Magnolia coco* (목련의 일종)의 101bp pair-end sequencing의 결과 data를 갖고 chloroplast 유전체의 de nove assemblly를 해 봅시다.

- 데이터의 생성:

Solexa system에서 생성된 약 10Gb의 data에서 reference sequence (이미 알고 있는 염기서열) 인 *Liriodendron tulipifera* (튤립나무)의 chloroplast whole genome (약 160kb)을 이용하여 BLAST search 하여 이것과 비슷한 sequence 들만 filtering 해 냄. 각 sequence 내에서 일정 quality 이상의 것만 사용하고 quality가 미달하는 것은 trimming.

- 각자의 번호에 해당하는 folder에 이렇게 마련된 데이터가 두 개의 file에 대응되는 pair-end sequece 들이 저장되어 있음 -> 100000-1.fastq 100000-2.fastq

TELNET program을 이용한 서버에 접속

1. telnet program 중 하나인 PuTTY를 web에서 찾아 인스톨
   - google에서 PuTTY
     http://www.chiark.greenend.org.uk/~sgtatham/putty/
   - Download PuTTY -> putty-0.62-installer.exe install
2. Open PuTTY
3. Host name에 www.amborella.net 입력 후 "open" click
4. 경고 메시지가 나오고 접속을 허용하면, 까만 화면의 Linux server에 접속화면이 나옴.

login as: lecture
Password: rkddml
[lecture@AMBORELLA ~]$ "$"가 나오면 접속 완료

Linux 기본 명령어들

$ ls	디렉토리와 파일 리스트 보기
$ ls -al	디렉토리와 파일 리스트 자세히 보기
$ cd M_kobus	M_kobus directory로 이동
$ cd ..	한단계 위의 디렉토리로 이동
$ pwd	자신이 작업하고 있는 prompt의 directory상의 위치가 어딘지 보여줌.
$ wc -l 100000-1.fastq	100000.02-1.fastq 파일내의 줄수 세기
$ rm 100000-1.fastq	100000.02-1.fastq화일을 삭제
$ rmdir K21	K21 폴더 전체를 삭제

VI editor로 파일 열어보기

Illumina data 가 들어있는 파일이 100000-1.fastq 라고 할 때

$ vi 100000-1.fastq 100000-1.fastq 파일을 vi editor로 열어봄

- 긴 파일이름 또는 디렉토리 등을 카피해서 쓰고싶으면 마우스로 카피할 곳을 하이라이트 시킨 후 Shift + Insert (PuTTY에서만 사용 가능)
- vi editor 속에서의 page up and down: Ctrl + F and Ctrl + B
- Fastq file은 4가지의 다른 줄이 반복됨.
- 1열: @"한 read의 고유이름", 2열: sequence,
  3열: +"한 read의 고유이름", 4열: sequence quality.
- sequence quality는 아스키II code의 기호의 숫자-66로 알 수 있다.
- vi editor에서 빠져나오기

esc를 누른 후 :q

VELVET program을 이용한 assembly

아래의 두 단계 과정을 거쳐 assembly함

$~toolbox/bin/velvet_1.2.01/velveth k61 61 -shortPaired -fastq 100000-1.fastq 100000-2.fastq

$~toolbox/bin/velvet_1.2.01/velvetg k61 -ins_length 500 -cov_cutoff auto -exp_cov auto

61mer option을 사용하여 k61 folder를 만들어 그 속에 결과 저장

- 여러개의 file이 k31에 생성되어 있는데, contigs.fa에 결과 있음.

$perl ~/stat_fasta.pl contigs.fa

여러 가지 statistics 나옴.

서버와의 접속 끊기 (SSH)

$logout

WinSCP program으로 server의 결과를 자신의 컴으로 갖고오기

1. google.co.kr에서 WinSCP를 검색
2. http://winscp.net/eng/docs/lang:ko 에서 WinSCP down 받기
3. WinSCP 설치 --> WinSCP download page --> downloads
4. WinSCP 실행
   호스트이름: www.amborella.net
   사용자이름: lecture
   password: rkddml
   위의 사항 저장 후 로그온
5. M-kobus folder 속의 자신의 번호 folder속의 특정 결과 folder (예: k67) 속의 모든 파일을 자신의 컴으로 복사: 파일을 모두 highlight한 후 drag 하면 됨.
6. WinSCP 종료

SEQUENCHER program을 이용한 Genome의 cover area의 확인

1. Sequencher를 다운받아 install (web site에 있음)
2. 알고 있는 reference sequence인 *Liriodendron*의 chloroplast whole sequence를 다운받아 저장 후 압축 풀기 (web site에 link 있음)
3. Sequencher open
4. Open sequence에서 *Liriodendron* whole sequence file을 open.
5. 서버에서 다운받은 자신의 contig 파일을 오픈
6. contig 들 전체와 "Liri_whole-IR" fragment를 지정하여 assemble.

여러 가지 조건으로 assemble한 것들 중 가장 contig수가 적고 각각의 contig의 길이가 긴 것을 선택하여 sequencher로 *Liriodendron*과 assemble 하여 assemble 한 것이 reference sequencer인 Liriodendron의 몇 %를 cover 했는지 파악.