

# NGS 데이터분석

## 3강 Fastq 포맷 및 다운로드

출처 : <http://www.tinyurl.com/ngs2016>

# NGS 데이터

- NGS 시퀀싱 장비에서는 각 제조사마다 독특한 형식의 base call 파일을 생성
- 대부분의 분석 소프트웨어들이 공통으로 받아들이는 표준 서열 파일 형식은 fastq 포맷
- 시퀀싱 장비에 부착된 소프트웨어들은 서열 데이터를 fastq 형식으로 변환하는 프로그램이 존재

# FASTA format

FASTA is pronounced "fast A", and stands for "FAST-All", because it works with any alphabet, an extension of the original "FAST-P" (protein) and "FAST-N" (nucleotide) alignment tools.

## An example of a multiple sequence FASTA file

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAK
KADRLAAEGVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKE
NEERRRLKDPNKPEHKQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNII
PGKMNSFIADNSQLDSKLTLMGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEF
ICFEVGEGLEKKTEDFAAEVAAQL

>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEE
YLKSQIATIGENLVVRRFATLKAGANGVVNGYIHTNGRVRGVVIAAACDSAEVA
SKSRDLLRQICMH
```

Fasta sequence always start with this sign

Fasta sequence will always have a sequence header

# FASTAQ format

- Next-Generation Sequencing(NGS) data를 표현할 때 쓰는 표준 format
- 일반적으로 fasta 형식의 파일은 염기 혹은 아미노산 서열을 표현할 때 사용하는데, 염기의 quality는 서열과 함께 저장하지 않고 별도의 파일에 저장
- Fastq는 서열과 quality를 한 파일에 저장하기 위해 고안된 형식

# FASTAQ format

- 하나의 read에 대한 정보는 총 4줄
- 첫번째줄
  - ✓ Header 로 @로 시작
  - ✓ 등록번호 + 시퀀스장비명 + "flowcell ID:tile number:spotX좌표:spotY좌표" + sequence 길이 로 구성
- 두번째줄 : sequence data
- 세번째줄 : +로 시작하고 첫번째 줄과 동일
- 네번째줄 : sequence의 quality에 대한 정보

```
@DRR000615.149 HWUSI-EAS505:1:1:15:14 length=51
GTAAGGGCACAACGTTTCTCTCAAGGGCCANNNNNTNNNNNNNTNNNNNNN
+DRR000615.149 HWUSI-EAS505:1:1:15:14 length=51
95?A/3@C@CC@A+ABCCBCCCC@#####!!!!!!#!!!!!!#!!!!!!
@DRR000615.1395 HWUSI-EAS505:1:1:123:13 length=51
CGACGACTGCCCGTGAGCGTGTCTCAGTCCGNNNNNNNNNNNNNNNGNNNNNNN
+DRR000615.1395 HWUSI-EAS505:1:1:123:13 length=51
@B(92.(==9>2@=@#####!!!!!!#!!!!!!
@DRR000615.3018 HWUSI-EAS505:1:1:221:15 length=51
TAGGAACACTTTCTCTATTATTCTGCCTATCANNNNNNNNNCNNNNNNN
+DRR000615.3018 HWUSI-EAS505:1:1:221:15 length=51
7AA5->=9@75CA3>BB6BB;BA9;;5=#####!
@DRR000615.3021 HWUSI-EAS505:1:1:221:13 length=51
GTAAAAGTACATCCNNNNNNNNNNNNNNNNNGNNNNNNNN
+DRR000615.3021 HWUSI-EAS505:1:1:221:13 length=51
:,B*?4?3;BBB@A@BCCBCB#####!!!!!!#!!!!!!
```

1 read data

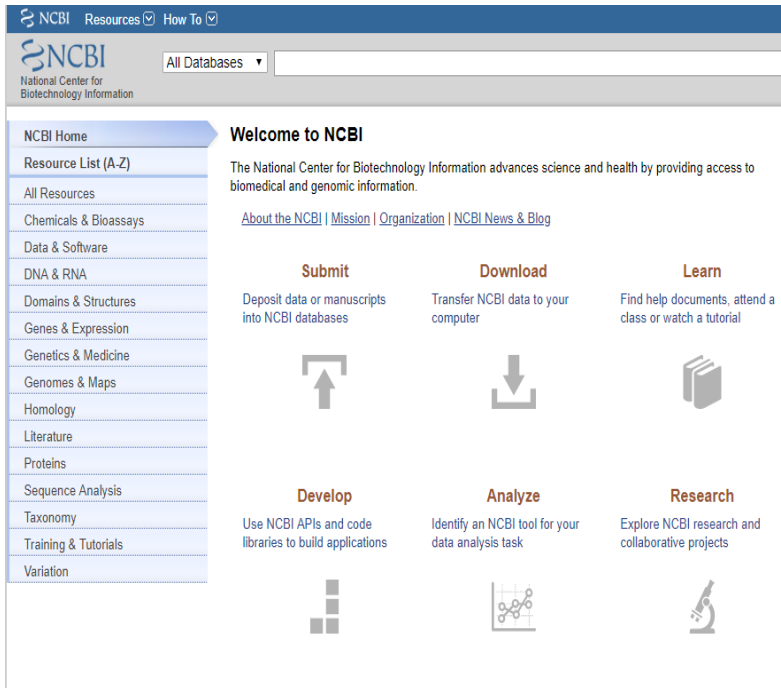
sequence data

# 등록번호

- NGS data를 사용한 연구 결과를 논문으로 출판하기 위해서는 NCBI/EBI/DDBJ 등과 같은 공공 데이터베이스에 등록
- NCBI는 SRR, EBI는 ERR, DDBJ는 DRR로 시작하는 등록번호를 부여
- "DRR00015"라는 NGS sequencing run의 데이터이며, 그 뒤에 따라오는 "1395"는 spot의 일련번호
- 일루미나 장비에서 직접 얻은 fastq 파일에는 위와 같은 등록번호없음. 예) @HWUSI-EAS100R:6:73:941:1973#0/1

# Fastq 파일 다운로드

- <https://www.ncbi.nlm.nih.gov/>
- All Databases -> SRA로 선택
- SRR490124 입력
- SRA Toolkit"이라는 소프트웨어 중에서 "fastq-dump"를 실행하여 다운로드할 것을 권장



NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

**Submit**  
Deposit data or manuscripts into NCBI databases

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class or watch a tutorial

**Develop**  
Use NCBI APIs and code libraries to build applications

**Analyze**  
Identify an NCBI tool for your data analysis task

**Research**  
Explore NCBI research and collaborative projects



SRA

SRR490124

Create alert Advanced

Full

**SRX143960: Escherichia coli K-12 isolate MT203**  
1 ILLUMINA (Illumina HiSeq 2000) run: 10.8M spots, 2.2G bases, 1.3Gb downloads

**Design:** Illumina Std PE Unamplified

**Submitted by:** DOE JOINT GENOME INSTITUTE (JGI)

**Study:** Resequencing of Escherichia coli K12 isolate MT203  
• [SRP012426](#) • [All experiments](#) • [All runs](#)  
[show Abstract](#)

**Sample:** Escherichia coli K-12 isolate MT203  
[SAMN00862350](#) • [SRS310604](#) • [All experiments](#) • [All runs](#)  
**Organism:** Escherichia coli str. K-12 substr. MG1655

**Library:**  
**Name:** HZSF  
**Instrument:** Illumina HiSeq 2000  
**Strategy:** WGS  
**Source:** GENOMIC  
**Selection:** RANDOM  
**Layout:** PAIRED

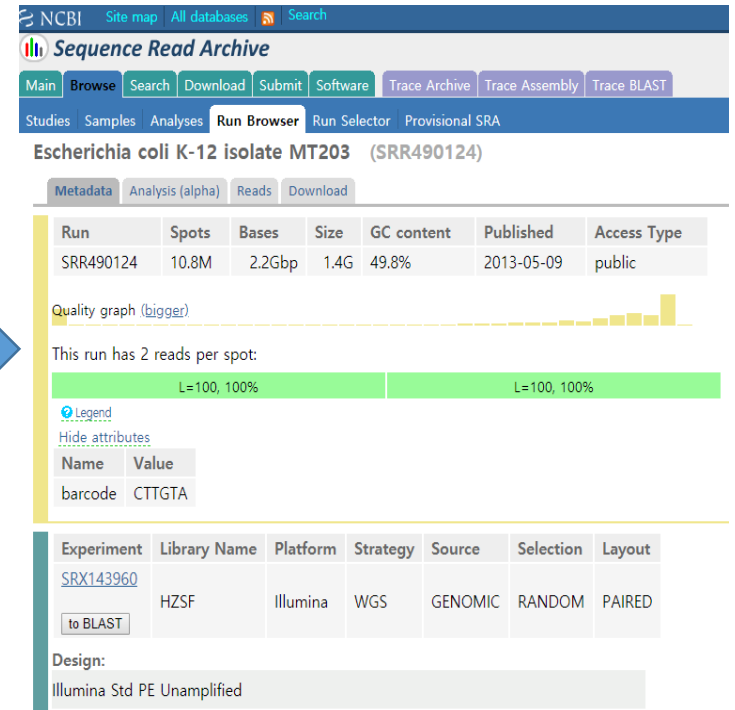
**Spot descriptor:**

1 forward 101 reverse

**Runs:** 1 run, 10.8M spots, 2.2G bases, 1.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR490124	10,776,679	2.2G	1.3Gb	2013-05-09

ID: 173995



NCBI Site map All databases Search

**Sequence Read Archive**

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

**Escherichia coli K-12 isolate MT203 (SRR490124)**

Metadata Analysis (alpha) Reads Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR490124	10.8M	2.2Gbp	1.4G	49.8%	2013-05-09	public

Quality graph (bigger)

This run has 2 reads per spot:

L=100, 100%

Legend

[Hide attributes](#)

Name	Value
barcode	CTTGTA

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
<a href="#">SRX143960</a>	HZSF	Illumina	WGS	GENOMIC	RANDOM	PAIRED

[to BLAST](#)

**Design:**  
Illumina Std PE Unamplified

# Fastq 파일 다운로드

fastq-dump -X 5 -Z SRR490124

```
명령 프롬프트
D:\DOWNLOAD\sratoolkit.2.9.0-win64\bin>fastq-dump -X 5 -Z SRR490124
Read 5 spots for SRR490124
Written 5 spots for SRR490124
@SRR490124.1 1706:3:1:1218:2186 length=200
CCCGGCATGGCGATAAAACGCCACCATCATGATGGACGTCGGGAACACAGTGGGGCGAGTTGGGTTCCACCTTGCAGAGATCAACGGAGAGCGTGGTAGAGCAAACAATCACAGCATG
TATTAATTGCCCTGCCACCCGCTGCTTCCCTGGACAGTTAGGCTTAGTCTCTGGTATCTTTGATATTTTCACGCTT
+SRR490124.1 1706:3:1:1218:2186 length=200
GGGGGGFEEGGAGGEGE7GGBGFFGFGGE: D?D; @C@?CCCC+@?@CB; D@@2D#####H?H; GHGHBHH>FHDE>@BE
3>D>45:8>; 2<75@C7A#####
@SRR490124.2 1706:3:1:1333:2075 length=200
ACCGACGCTTATCGCAGATTAGCACGTCCTTCATCGCCTCTGACTGCCAGGGCATCCACCGTGTACCGTTAGTCGCTTAAACTCAGCAACACAAGATTGTAATCTGGATCAAGCTGAAA
ATTGAAACACTGAACAATGAAAGTTGTTCTGAGTCTCTCAAAATTTTCGCCACCAAGATGATGAATOGCAAAAAAOCGCTT
+SRR490124.2 1706:3:1:1333:2075 length=200
HHFHHEFH=GFDFDHBEHFCEHHCHFHEAF<?>9>C>AA9@@C+5C1;; 2?5?<?#####BHFHFHHEBFFFE@EBECE
F@EGGEEEEHBECAEEBECDA?A5>4<; 5<972) '%+/@B9?@B#####
@SRR490124.3 1706:3:1:1329:2094 length=200
AAATAACAGTAATGATCGCTTTCTGACGCAGCATTTTCAGCGCTGGCTAGCGCAACGGCATTGCGCAAAACCTGATTACCTGCACGCGAGAAAAAATGCTTAAACTGCTGATTATGTC
GGTGAACCACTGAAAAAACACGTGAAGCAAAACGCTACCACAATATCTGCCACAGCATTACACGGAGCTCCAGGTCGTG
+SRR490124.3 1706:3:1:1329:2094 length=200
FFFFFFEFBFFDFEFEFF?4;>@C@9,>>,<<.; 7DDB(A#####GEFBEE@#####
#####
@SRR490124.4 1706:3:1:1378:2178 length=200
GTGGTTTGGCCACCGGATCCATGATTGCTGGTTGATAACAAAATCACTACACTAACGCCGTTCTCAACGGGGAGCCACGCCCTACGCGCTCCAGTGAAAAATTATCCGGATCAGGTTGAC
GGGTATTTTCTCAGCGCACGCGTACGCGTGGCACCCCGTTGAGAACGTGGTTAGTTGGGTGCTTTTGTGTGCAGACGCAC
+SRR490124.4 1706:3:1:1378:2178 length=200
GFGGGGBGFGGABGBG?FFD5BDD=FCC; CDC>=<?DC<6<1+3=????EE9@FE9AD#####HHGHGHHH@; @BE.58; 67
GEE. E<46<01&; 7A#####
@SRR490124.5 1706:3:1:1261:2194 length=200
AACGCCACCAAGCATCATTGCGGCTGTTGTGGCGTCAGGATCCCTGCCTGATGCTCACCATAAACGGAGTATTATCTCTTCGGTCGCACGGTGACCGCGGTTTACTGAAAGTGATAGGT
GCACTGGCGCTGGCGGGGAGGTGTCCGGCCGCTCATGACCAAAAAAGTAACATACGGCGCGCAACGCTGCGACCAAGACGC
+SRR490124.5 1706:3:1:1261:2194 length=200
HHHFFGHGHAHFGHGGFEGCCDGBAADAA?+B#####C@A?+A@>7@@:;; C16'43
5(714A7,C#####
```

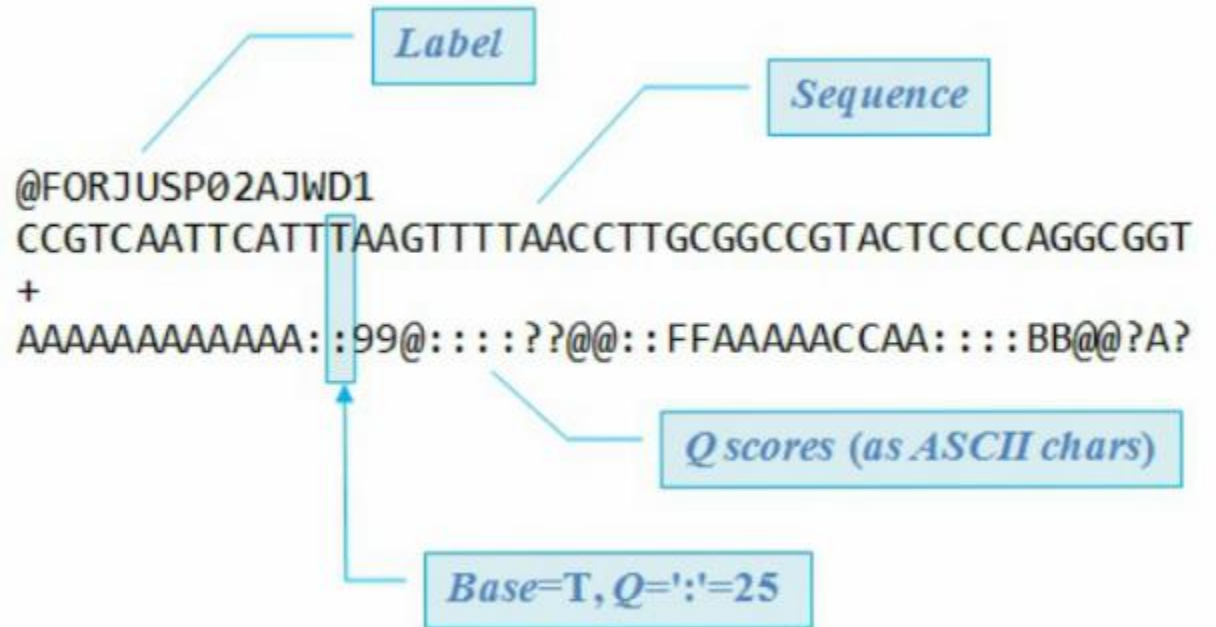


# Base Call Quality

- Sequence quality는 phred score를 나타냄

$$Q = -10\log_{10}P$$

- P는 염기가 틀릴 확률
- 'A'가 아닐 확률이  $10^{-2}$ 라면, phred score는 20
- phred score가 크면 클수록 맞을 확률이 높고 틀릴 확률이 적은 것, 0 ~ 40사이에 존재
- Phred score를 그대로 쓰면 자리가 맞지 않음
- phred score를 표현하기 위해 ASCII code를 사용



# ASCII code

- computer의 글자를 8개의 bit로 저장하는 규약이다 (256개의 글자를 저장할 수 있다)
- Phred score에 +33을 해준 값을 ASCII code로 바꿔서 fastq 파일에 저장
- Phred score 0 이면 !로 표시
- Phred score 40 이면 I로 표시
- A의 Phred score는 얼마인가 ??

ASCII control characters			ASCII printable characters			Extended ASCII characters										
00	NULL	(Null character)	32	space	64	@	96	`	128	Ç	160	á	192	À	224	Ó
01	SOH	(Start of Header)	33	!	65	A	97	a	129	ü	161	í	193	Á	225	Ô
02	STX	(Start of Text)	34	"	66	B	98	b	130	é	162	ó	194	Â	226	Õ
03	ETX	(End of Text)	35	#	67	C	99	c	131	â	163	ú	195	Ã	227	Ö
04	EOT	(End of Trans.)	36	\$	68	D	100	d	132	ä	164	ñ	196	ä	228	ö
05	ENQ	(Enquiry)	37	%	69	E	101	e	133	å	165	Ñ	197	å	229	Û
06	ACK	(Acknowledgement)	38	&	70	F	102	f	134	ä	166	ª	198	ä	230	µ
07	BEL	(Bell)	39	'	71	G	103	g	135	ç	167	º	199	Å	231	þ
08	BS	(Backspace)	40	(	72	H	104	h	136	ê	168	¿	200	⌘	232	ð
09	HT	(Horizontal Tab)	41	)	73	I	105	i	137	ë	169	®	201	⌘	233	ú
10	LF	(Line feed)	42	*	74	J	106	j	138	è	170	¬	202	⌘	234	û
11	VT	(Vertical Tab)	43	+	75	K	107	k	139	ï	171	½	203	⌘	235	ü
12	FF	(Form feed)	44	,	76	L	108	l	140	î	172	¾	204	⌘	236	ý
13	CR	(Carriage return)	45	-	77	M	109	m	141	ï	173	¿	205	⌘	237	ÿ
14	SO	(Shift Out)	46	.	78	N	110	n	142	Ä	174	«	206	⌘	238	
15	SI	(Shift In)	47	/	79	O	111	o	143	Å	175	»	207	⌘	239	'
16	DLE	(Data link escape)	48	0	80	P	112	p	144	É	176	⌘	208	ð	240	≡
17	DC1	(Device control 1)	49	1	81	Q	113	q	145	æ	177	⌘	209	Ð	241	±
18	DC2	(Device control 2)	50	2	82	R	114	r	146	Æ	178	⌘	210	É	242	
19	DC3	(Device control 3)	51	3	83	S	115	s	147	ò	179	⌘	211	Ê	243	¼
20	DC4	(Device control 4)	52	4	84	T	116	t	148	ó	180	⌘	212	Ë	244	½
21	NAK	(Negative acknowl.)	53	5	85	U	117	u	149	ô	181	À	213	Ì	245	¾
22	SYN	(Synchronous idle)	54	6	86	V	118	v	150	ù	182	Á	214	Í	246	÷
23	ETB	(End of trans. block)	55	7	87	W	119	w	151	ù	183	Â	215	Î	247	°
24	CAN	(Cancel)	56	8	88	X	120	x	152	ý	184	©	216	Ï	248	´
25	EM	(End of medium)	57	9	89	Y	121	y	153	Û	185	⌘	217	⌘	249	ˆ
26	SUB	(Substitute)	58	:	90	Z	122	z	154	Ü	186	⌘	218	⌘	250	˜
27	ESC	(Escape)	59	;	91	[	123	{	155	ø	187	⌘	219	⌘	251	¸
28	FS	(File separator)	60	<	92	\	124		156	£	188	⌘	220	⌘	252	˙
29	GS	(Group separator)	61	=	93	]	125	}	157	Ø	189	¢	221	⌘	253	˚
30	RS	(Record separator)	62	>	94	^	126	~	158	×	190	¥	222	⌘	254	■
31	US	(Unit separator)	63	?	95	_			159	f	191	Ÿ	223	⌘	255	nbsp
127	DEL	(Delete)														