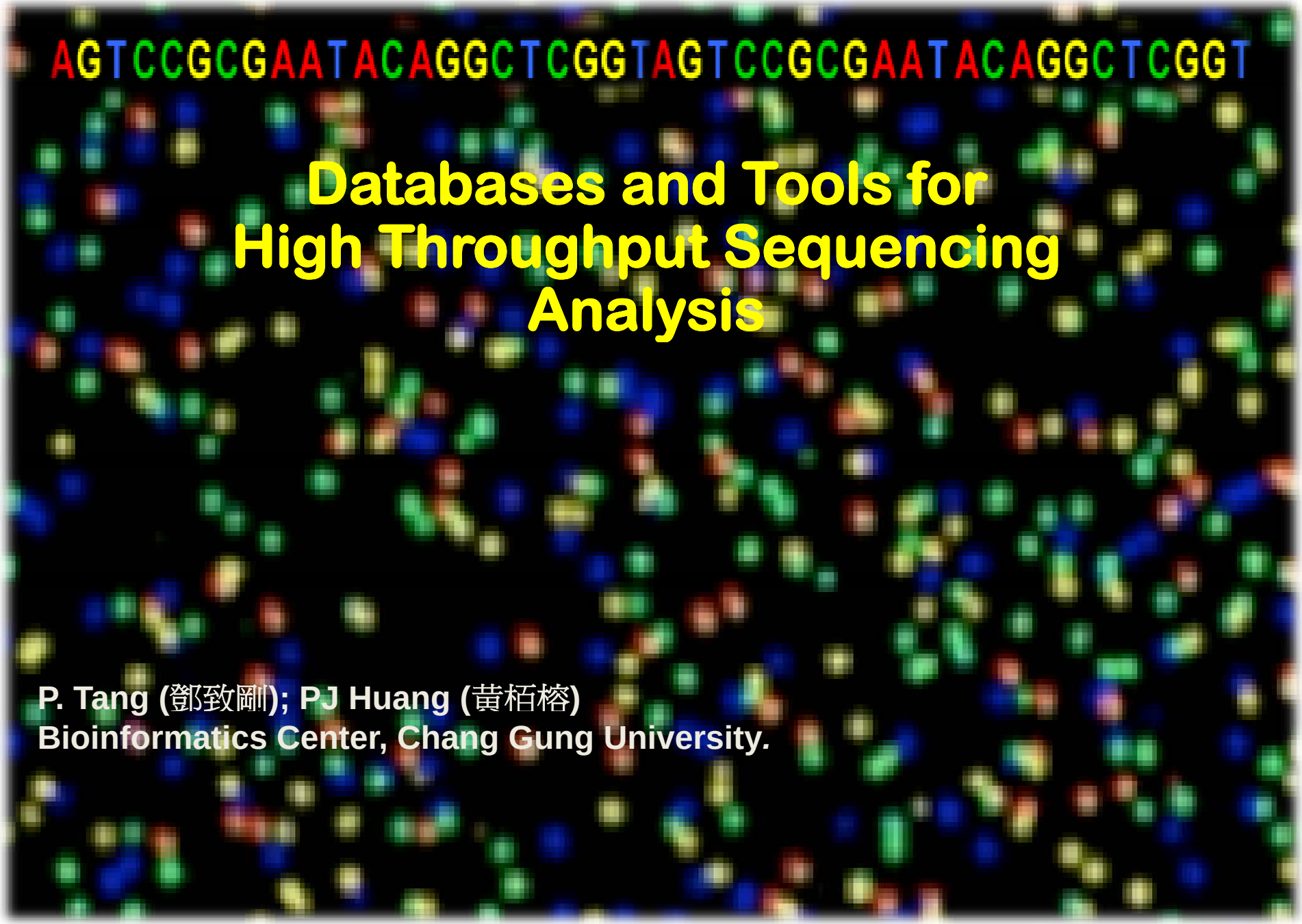


Databases and Tools for High Throughput Sequencing Analysis

P. Tang (鄧致剛); PJ Huang (黃栢榕)
Bioinformatics Center, Chang Gung University.

上課時間：每週五上午 08:10~12:00 （08:10~09:00—授課，09:10~12:00—實習）
上課地點：醫學院 9F 生物資訊中心 電腦教室

助教：方怡凱(分機5690) ，鄭尉弘(分機5690)

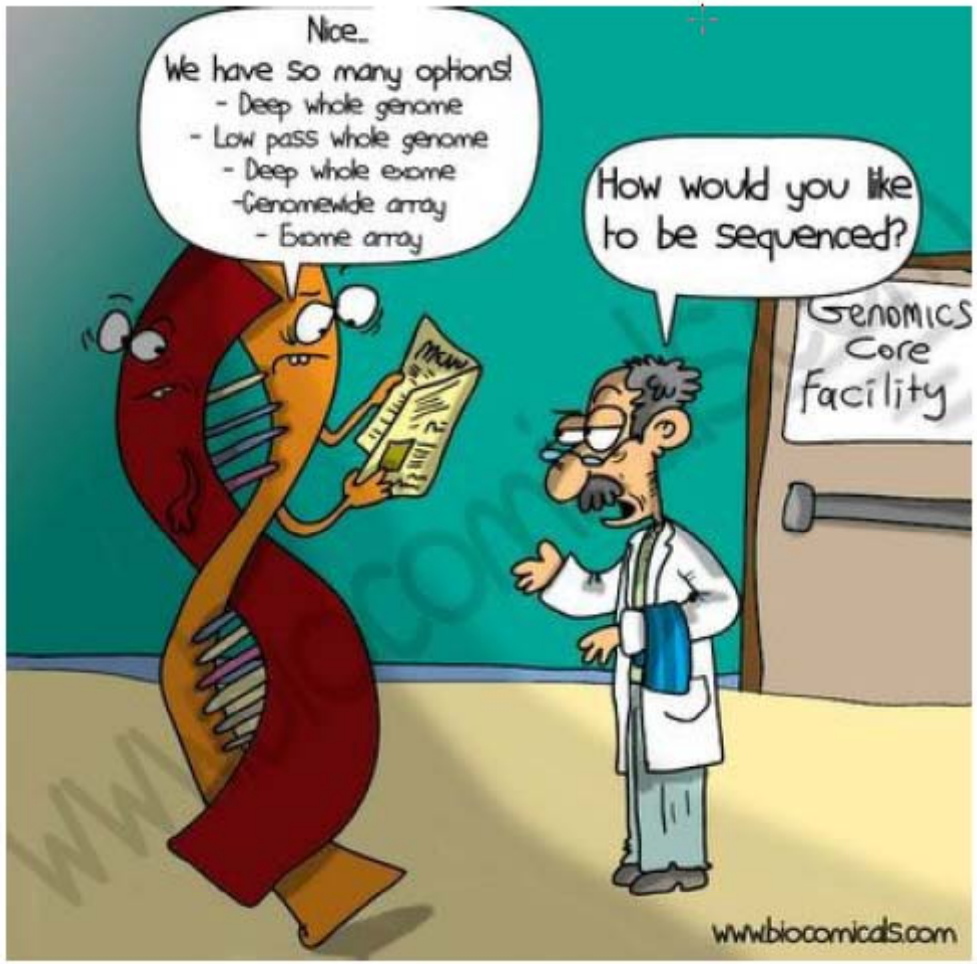
進度	課程進度	授課老師	講義下載 L:lecture T:tutorial E:exercise
1	HOLIDAY		
2	Introduction to New Generation Sequencing (NGS) Technologies	鄧致剛	L
3	Databases and Tools for High Throughput Sequencing Analysis	黃柏榕	L T:FastQC **Movie:**FastQC T:FASTX T:SRA
4	Small RNA Analysis (I) and Target Analysis	黃柏榕	L T:CLC DSAP miRDEEP E
5	RNA Sequencing (De novo) Analysis	鄧致剛	Lecture Exercise Unix T1 T2
6	HOLIDAY		
7	Transcriptome Functional Profiling-1 (Metacore & DAVID)	鄧致剛	
8	Transcriptome Functional Profiling-2 (GSEA & KEGG)	鄧致剛	
9	MID-TERM EXAMINATION	鄧致剛	
10	Genome Resequencing (exercise)	鄧致剛 李季青	Lecture Exercise T
11	Small Genome de novo Assembly	鄧致剛 陳亭妏	
12	Bacteria Genome & Gene Expression Analysis	陳怡原	Lecture
13	Genome Annotation & Genome Browser	鄧致剛 李季青	Lecture+Exercise
14	Metagenome Analysis	鄧致剛 甘瑞麒	
15	Chip Sequencing Analysis	鄧致剛 陳亭妏	Lecture+Exercise
16	Targeted Resequencing Analysis	黃柏榕	
17	Student Presentation	鄧致剛	PDF
18	Final-Examination	鄧致剛	PDF

HTseq Platforms

- 454 Sequencing / Roche
  - GS Junior System
  - GS FLX+ System
- Illumina (Solexa)
  - HiSeq System
  - Genome analyzer IIx
  - MySeq
- Applied Biosystems - Life Technologies
  - SOLiD 5500 System
  - SOLiD 5500xl System
- Ion Torrent - Life Technologies
  - Personal Genome Machine (PGM)
  - Proton
- Helicos
  - Helicos Genetic Analysis System
- Pacific Biosciences
  - PacBio RS
- Oxford Nanopore Technologies
  - GridION System
  - MinION

Next Generation Sequencing
Amplified Single Molecule Sequencing

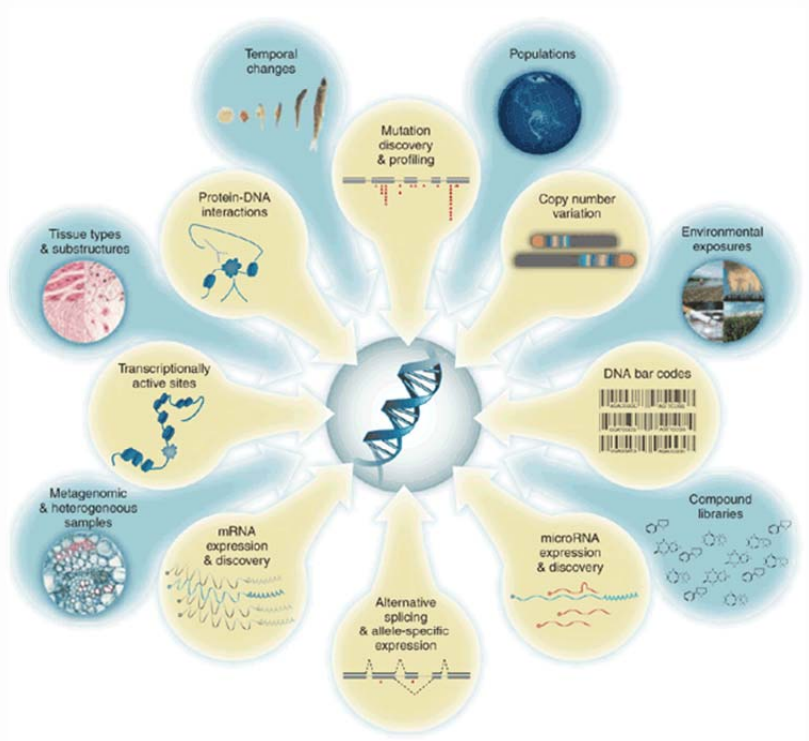
Third Generation Sequencing,
Next Next Generation Sequencing,
Single Molecule Sequencing



- Genome Sequencing
  - De novo sequencing
  - Resequencing
  - Targeted (re)sequencing
  - Mitochondrial sequencing
  - Mutation detection
  - Amplicon sequencing
  - Amplicon Cancer Panel
  - ...
- Transcript Expression Profiling
  - RNA sequencing
  - miRNA sequencing
  - Deep-SAGE
  - Deep-CAGE
  - PAS : polyadenylation site
  - ...
- Transcription factor binding
  - ChIP sequencing
- Structural variation
- Metagenomics
- ...

Applications on Biomedical Sciences





DNA
Exome Sequencing
Genome Resequencing
De novo Genome Sequencing
Meta-Genomic Sequencing
Chip-Sequencing

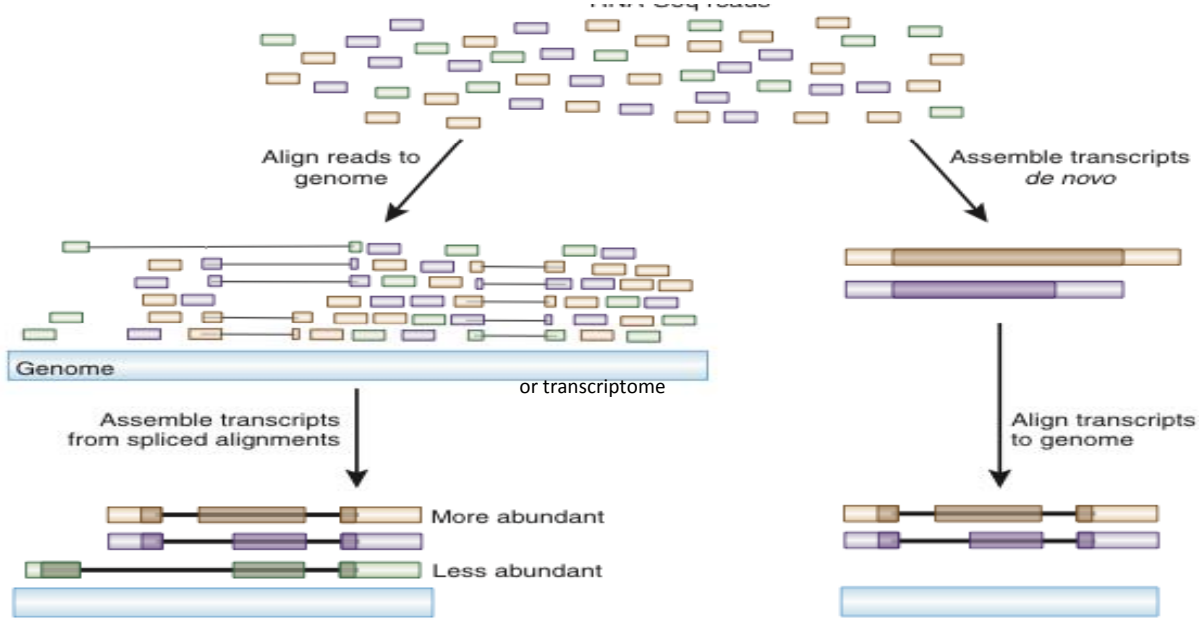
RNA
Small RNA Sequencing
Transcriptome Sequencing
De novo Transcriptome Sequencing
Meta-Transcriptome Sequencing

BIOINFORMATICS



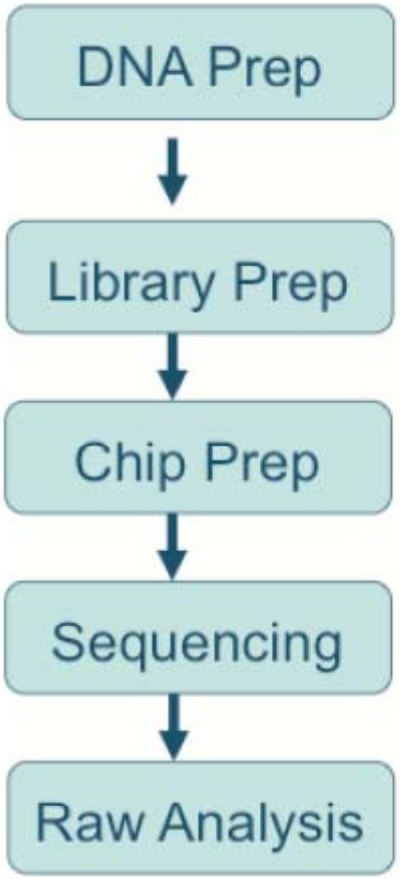
IBM X3850

Analysis Strategies: Reference Sequence *Alignment* (Mapping) *vs* De novo *Assembly*



Process	Software & Algorithms	Website
Preprocessing step	homemade script	(N/A)
(1.1) Alignment	MAQ	http://maq.sourceforge.net
	BWA	http://bio-bwa.sourceforge.net/bwa.shtml
	BWA-SW (SE only)	http://bio-bwa.sourceforge.net/bwa.shtml
	PERM	http://code.google.com/p/perm/
	BOWTIE	http://bowtie-bio.sourceforge.net
	SOAPv2	http://soap.genomics.org.cn
	MOSAIK	http://bioinformatics.bc.edu/marthlab/Mosaik
	NOVOALIGN	http://www.novocraft.com/
(1.2) *De novo* Assembly	VELVET	http://www.ebi.ac.uk/%7Ezerbino/velvet
	SOAPdenovo	http://soap.genomics.org.cn
	ABYSS	http://www.bcgsc.ca/platform/bioinfo/software/abyss

HTseq Experiment



Randomly shear DNA + end repair + size select

Append sequencing adapters

Layout of library on sequencing slide or wells

For each library fragment – determine the order and identity of bases at either end of the fragment

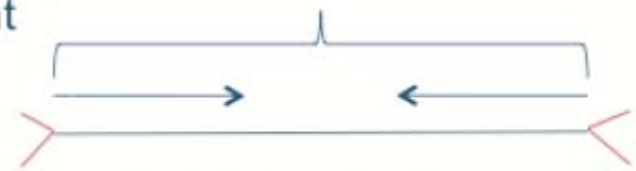


Image processing + base calling

- Base calls + associated quality: Fastq/BAM

Great… I got my data now what…

- Data and information management is slowly moving out of infancy in genomics science…. at the toddler stage…

- **The Good news**
  - Some data formats are being accepted widely

- **The Bad news**
  - Still many competing standards in some areas
  - Interoperability of data standards is almost non-existent
  - Governance is questionable

Storage & Computing Power

Next gen sequencers generated Giga bp to Tera bp of data

Illumina HiSeq 2000	SOLiD 5500xl	454 FLX Titanium	Ion Torrent Personal Genome Machine	PacBio RS
100 bp reads	75bp x 35bp or 60bp x 60 bp	250-400 bp reads	200 bp reads	1000 bp reads
8 days	7 days	10 hours	2 hours	< 1 hour



TABLE 1. Average file sizes for various analyses

	Image Data Size†	Primary Analysis Data*	Secondary Analysis Data
1 slide - 1 tag (fragment library)	1.8 - 2.4 TB	< 100 GB	10-20 GB
1 slide - 2 tags (mate-paired library)	3.5 – 4.0 TB	< 100 GB	30-40 GB
2 slides - 1 tags (fragment library)	5.0 TB	< 100 GB	20-40 GB
2 slides - 2 tags (mate-paired library)	8.2 TB	< 100 GB	60-80 GB

Data Format Types

- Raw Sequence Data e.g. fasta

```
>xyz some other comment
ttcctctttctcgactccatcttcgcggtagctgggaccgccgttcagtcgccaatatgc
agctctttgtccgcgcccaggagctacacaccttcgaggtgaccggccaggaaacggtcg
cccagatcaaggctcatgtagcctcactggagggcatt
```

- Aligned data e.g. BAM
  - SAM (Sequence Alignment/Map) format has become the *de facto* standard for storing alignment data.
  - BAM is a binary version of SAM allowing more efficient storage.
- Processed data e.g. BED

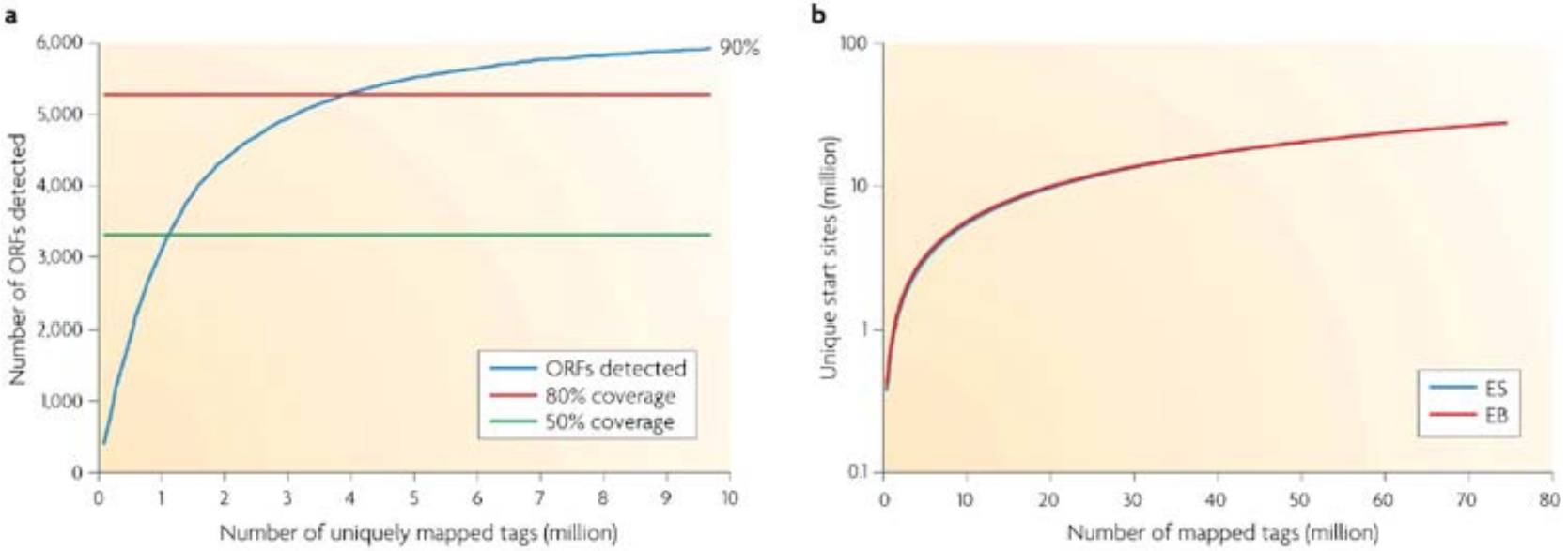
SAM format

```
ERR005646.11088674      147     1       161099954       60
54M     =       161099742       -266
TTTTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACT
89=<;@>EECFCBBFFCAEFBGB=FFFC?@AB@G=FFB@CABABA?A@<>>=;=
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646  OQ:Z:D?
FFEEEFFFFFFFFFFEFFDFECFFFE;EEEEFCFFEEEEFEFECEEC=E;EF
ERR005646.5518024       147     1       161099956       60
54M     =       161099847       -163
TTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACTCT  :
68=<A@@A???AB?A>ABBB>@CABCAAA>B@BAB@BA@A@A@A@=A=A=>;<
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646
OQ:Z:ECEEEEEEEEDEEEEE>EEEEEEEEEEEEE@EEEEBEEEEEEEEECCCBEEEE
```

Interpreting raw data



How deep should we go?
coverage



(a) 80% of yeast genes (**genome size: ~120MB**) were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3' end.

(b) The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body.

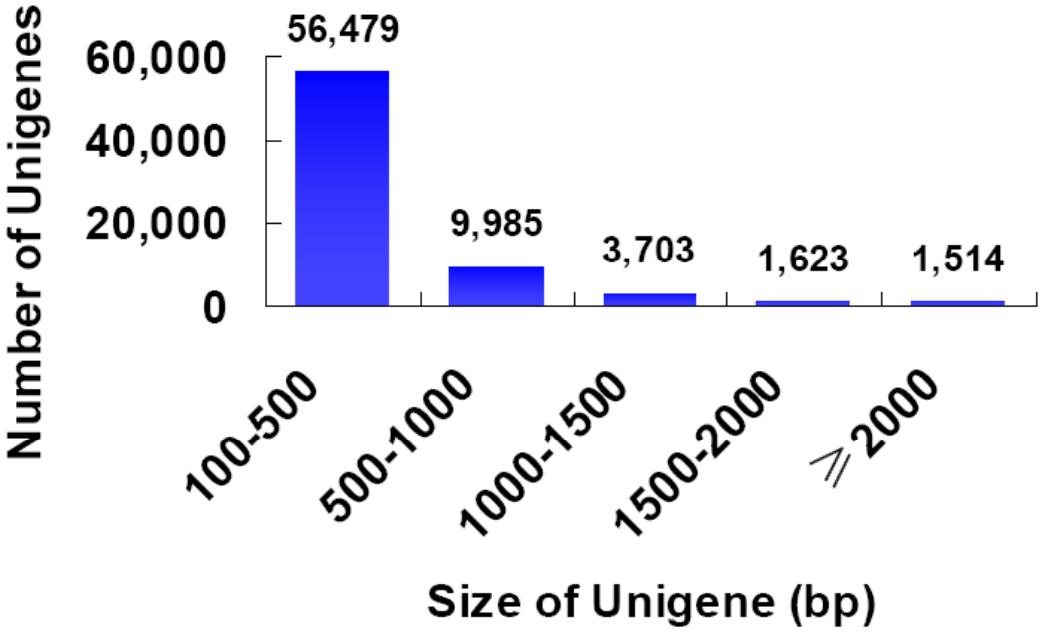
Nature Reviews Genetics ***10***, *57-63*

Genome Size

De novo assembled rice transcriptome

1.3 Gb RNA-Seq data (genome size: ~400MB)

85% of assembled unigenes were covered by gene models

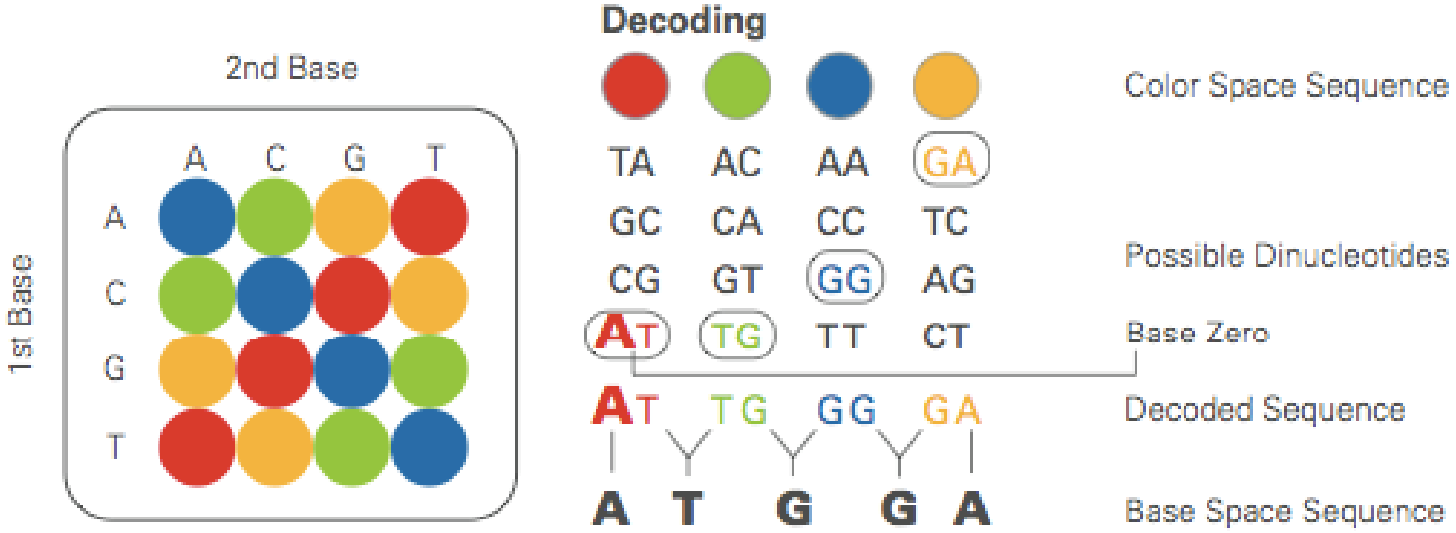


HTseq Raw Data Format

- fasta (Sanger)
- csfasta (SOLiD)
- fastq (Solexa)
- sff (454)
- …. And about 30 other file formats

- http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

SOLiD Color Space



Its format is

```
>TAG_ID
Color_space
```

```
>1_88_1830_R3
G3211312320130023232O
>1_89_1562_R3
G2313313123333310132O
```

(cs)Fasta/(cs)Fastq

- FASTA
  - Header line “>”
  - Sequence
- FASTQ
  - Add QVs encoded as single byte ASCII codes

```
@NA12878:1463:NA12892:NA12891:F_IL20_290:1:80:114:644
TTTGCATTTAACAAATAATATGAGAACCGTTGACTG
+
6@<?3@@5@7@AAABB1A;;;BBABABB<@==<9/.
@NA12878:1463:NA12892:NA12891:F_IL20_290:3:97:342:584
GCATTTAACAAATAATATGAGAACCGTTGACTGAAA
+
@@AA@AAABAAABBABBABB>>BABAACA=@@A@<<
@NA12891:1463:::M_IL6_344:6:73:359:297.2
TTTCAGTCAACGGTTCTCATATTATTTGTTAAATGC
+
????>>??@?@@@AAA;A@AAA@:@@AA@@;4-4;:
```

- Most aligners accept FASTA/Q as input
- Issue: data is volumous (2 bytes per base for FASTQ)
- Do PHRED scaled values provide the most information?

Fastq: Illumina & Snager

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

```
@HWI-E4_9_30WAF:1:1:8:308
TCCACATCAGAGGCCATGGCCACCAGGCCCAGGAT
+HWI-E4_9_30WAF:1:1:8:308
aaaaXaaabaa^aaLaaLLa^a^^VV\aaaaaaaa
@HWI-E4_9_30WAF:1:1:9:947
CCAATGTGGTCATAGGTGACAACCTTCTCCTCGCT
+HWI-E4_9_30WAF:1:1:9:947
aZaaaaaaaZaab^aaaWaaaaaaaaaaaaa\aaa
@HWI-E4_9_30WAF:1:1:9:1505
GGAAGCCAGGACCCACCATGAGTAGCATACATCTG
F:1:1:9:1505
```

Fastq: Illumina & NCBI

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)*
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

sff (text format): 454

```
>F7K88GK01BMPI0
Run Prefix: R_2009_12_18_15_27_42_
Region #: 1
XY Location: 0551_2346

Run Name: R_2009_12_18_15_27_42_FLX########_Administrator_yourrunname
Analysis Name: D_2009_12_19_01_11_43_XX_fullProcessing
Full Path: /data/R_2009_12_18_15_27_42_FLX########_Administrator_yourrunname/
D_2009_12_19_01_11_43_XX_fullProcessing/

Read Header Len: 32
Name Length: 14
# of Bases: 500
Clip Qual Left: 15
Clip Qual Right: 490
Clip Adap Left: 0
Clip Adap Right: 0

Flowgram: 1.03 0.00 1.01 0.02 0.00 0.96 0.00 1.00 0.00 1.04 0.00 0.00 0.97 0.00 0.96 0.02
0.00 1.04 0.01 1.04 0.00 0.97 0.96 0.02 0.00 1.00 0.95 1.04 0.00 0.00 2.04 0.02 0.03 1.05
Flow Indexes: 1 3 6 8 10 13 15 18 20 22 23 26 27 28 31 31 34 35 37 37 37 40 43 45 47 47 47 50
53 53 53 55 58 60 63 66 67 67 67 67 70 71 71 74 74 76 79 82 83 86 86 88 88 91 93 96 97...
Bases:
tcagatcagacacgCCACTTTGCTCCCATTTCAGCACCCCACCAAGCACAAGGCTGTCATCCCAATTGGACGGACAGATATGAGGTTAGCATT
GGAAACCAATTCAGTCCCTAATTATTCACGACTGAACCCAGCGACAATTGGACATGGATTCATTTTTCAACTTGATTTGTTGTTGTAAAAGCA
...
Quality Scores: 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 38 38 38 40 40 40 39 39
```

454 fasta with quality file

Fasta:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
AGAAAGTCAGCGGCAAATTTGGTTTTAGACGAATTGTCCCTTTGACATAACGACTAAAGG
AGTCAACAGATTTTCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACGCTATT
...
```

Qual:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
40 40 39 39 39 40 40 40 40 40 40 40 40 38 31 26 26 16 16 16 20 20
14 14 14 14 27 33 32 35 36 33 36 35 36 38 35 20 20 21 24 24 22 36
39 40 38 38 38 40 40 40 40 40 40 37 37 37 33 33
29 36 38 38 38 38 38 38 38 35 20 21 21 21 31 36 37 40 40 35 37 37
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
...
```

Sanger-style Phred scores $Q_{\text{sanger}} = -10 \log_{10} p$

454 base quality?

Fasta:

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
AGAAAGTCAGCGGCAAATTTGGTTTTAGACGAATTGTCCCTTTGACATAACGACTAAAGG
AGTCAACAGATTTTCGTATAACTTCGTATAATGTATGCTATACGAAGTTATTACGCTATT
...
```

Qual:

chance of being wrong: $1:10^{4.0} = 1:10000$

```
>FTJD6BE02HHD3W length=409 xy=2951_1562 region=2 run=R_2009_04_01_11_28_49_
40 40 39 39 39 40 40 40 40 40 40 40 40 38 31 26 26 16 16 16 20 20
14 14 14 14 27 33 32 35 36 33 36 35 36 38 35 20 20 21 24 24 22 36
39 40 38 38 38 40 40 40 40 40 40 37 37 37 33 33
29 36 38 38 38 38 38 38 38 35 20 21 21 21 31 36 37 40 40 35 37 37
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
...
```

chance of being wrong: $1:10^{3.5} = 1:3162$

Sanger-style Phred scores

All Platforms have Errors



Illumina



SoLID/ABI-Life



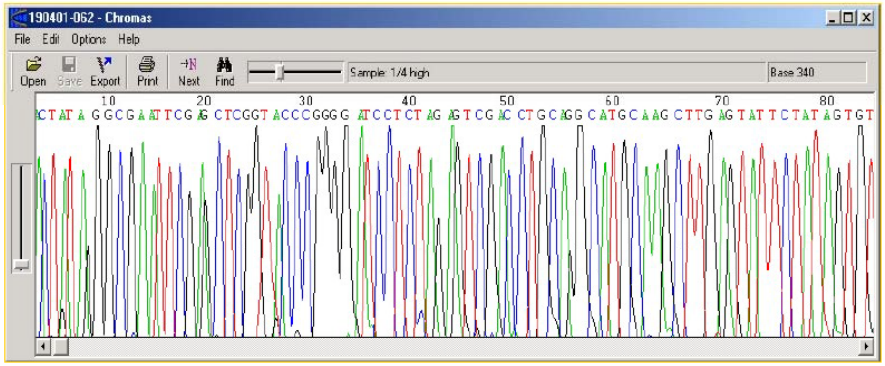
Roche 454



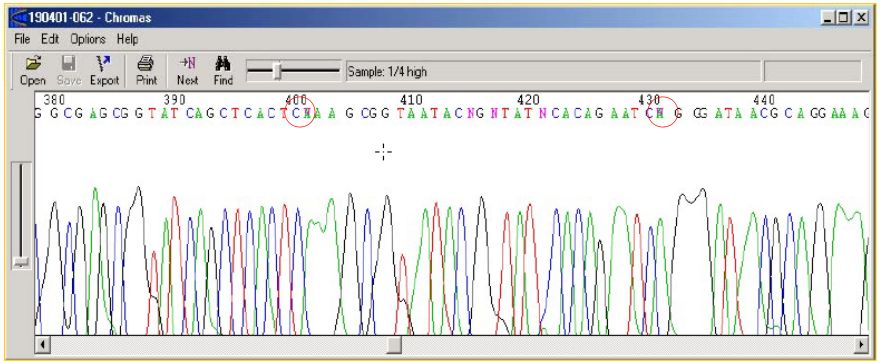
Ion Torrent

1. Removal of low quality bases/ Low complexity regions
2. Removal of adaptor sequences
3. Homopolymer-associated base call errors (3 or more identical DNA bases) causes higher number of (artificial) frameshifts

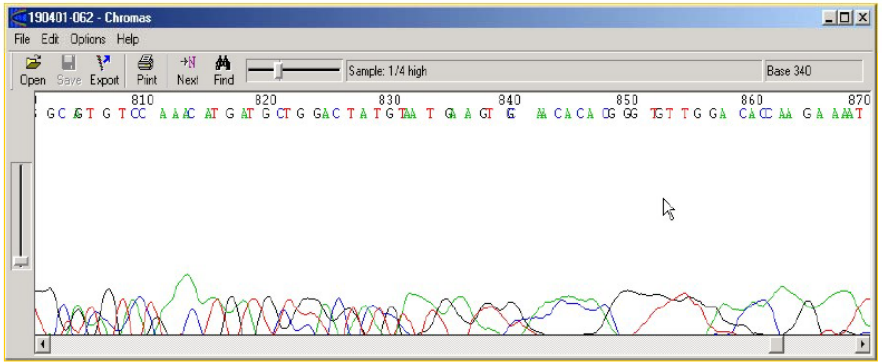
Trace File



High quality region - NO ambiguities (Ns)



Medium quality region - SOME ambiguities (Ns)



Poor quality region - LOW confidence

Quality Control Is Essential

Sequencing runs cost money

-Sequencing a poor library on multiple runs – throwing money away!

-Cost of analyzing data $

-CPU time required for alignment and assembly $$

-Cost of storing raw sequence data $$$

-Hours of analysis could be wasted $$$$

Two aspects of sequencing QC

Run/Lane QC

-Is the data off this lane useful?

Library QC

-Is this library worth sequencing more lanes off?

E.g. Have 2x but want another 8x in total

Slightly different set of metrics

Accessing Quality: phred scores

Phred quality scores were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each base call in automated sequencer traces.[1][2] Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences.

http://en.wikipedia.org/wiki/Phred_quality_score

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

P=error probability of a given base call

Genome Research **8**: 175-185, 1998

Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment

Brent Ewing,[1] LaDeana Hillier,[2] Michael C. Wendl,[2] and Phil Green[1,3]

[1]Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA;
[2]Genome Sequencing Center, Washington University School of Medicine, Saint Louis, Missouri 63108 USA

Genome Research **8**: 186-194, 1998

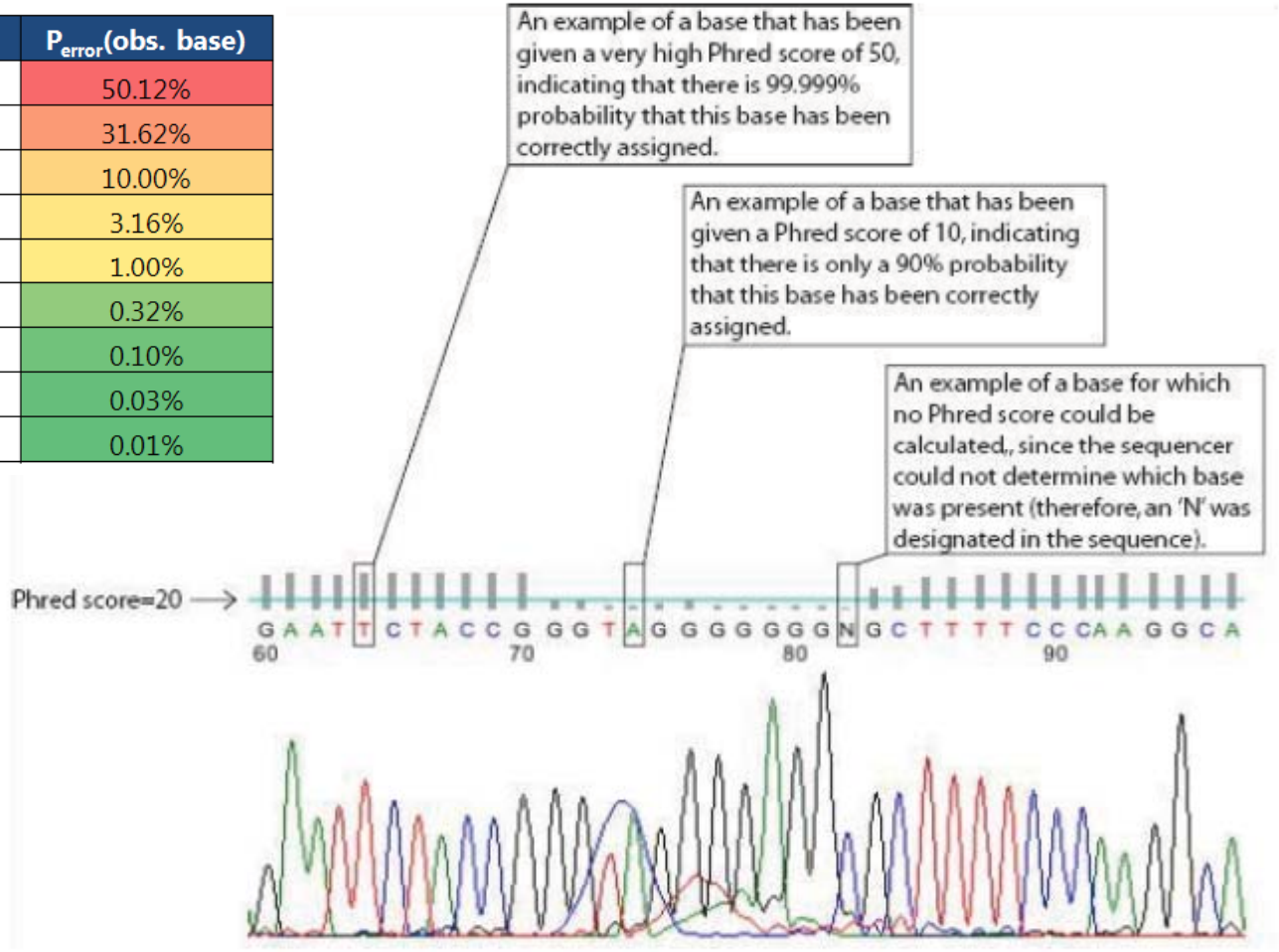
Base-Calling of Automated Sequencer Traces Using *Phred*. II. Error Probabilities

Brent Ewing and Phil Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195-7730 USA

Accessing Quality: phred scores

Base Quality	P_{error}(obs. base)
3	50.12%
5	31.62%
10	10.00%
15	3.16%
20	1.00%
25	0.32%
30	0.10%
35	0.03%
40	0.01%



http://en.wikipedia.org/wiki/Phred_quality_score

454 output formats

Standard flowgram *format*

Read name

.sff

Flowed nucleotide*

Flow no.

Flow value (light signal strength)

Quality scores

oedipus.imr.no - PuTTY

FR3GPDN01CVFQV	786	a	0.10	
FR3GPDN01CVFQV	787	c	0.09	
FR3GPDN01CVFQV	788	g	0.08	
FR3GPDN01CVFQV	789	t	0.10	
FR3GPDN01CVFQV	790	a	0.10	
FR3GPDN01CVFQV	791	c	0.09	
FR3GPDN01CVFQV	792	g	0.08	
FR3GPDN01CVFQV	793	t	0.10	
FR3GPDN01CVFQV	794	a	0.10	
FR3GPDN01CVFQV	795	c	0.09	
FR3GPDN01CVFQV	796	g	0.07	
FR3GPDN01CVFQV	797	t	0.14	
FR3GPDN01CVFQV	798	a	0.11	
FR3GPDN01CVFQV	799	c	0.08	
FR3GPDN01CVFQV	800	g	0.09	
FR3GPDN01A5TLK	1	t	0.97	37
FR3GPDN01A5TLK	2	a	0.00	
FR3GPDN01A5TLK	3	c	1.01	37
FR3GPDN01A5TLK	4	g	0.04	
FR3GPDN01A5TLK	5	t	0.02	
FR3GPDN01A5TLK	6	a	1.02	37
FR3GPDN01A5TLK	7	c	0.00	
FR3GPDN01A5TLK	8	g	2.10	37,37
FR3GPDN01A5TLK	9	T	2.91	37,37,37
FR3GPDN01A5TLK	10	A	0.01	
FR3GPDN01A5TLK	11	C	0.00	
FR3GPDN01A5TLK	12	G	1.01	37
FR3GPDN01A5TLK	13	T	2.15	37,37
FR3GPDN01A5TLK	14	A	0.00	
FR3GPDN01A5TLK	15	C	0.00	
FR3GPDN01A5TLK	16	G	1.01	37
FR3GPDN01A5TLK	17	T	0.36	
FR3GPDN01A5TLK	18	A	0.94	37
FR3GPDN01A5TLK	19	C	0.00	
FR3GPDN01A5TLK	20	G	1.01	37
FR3GPDN01A5TLK	21	T	0.10	
FR3GPDN01A5TLK	22	A	1.07	37
FR3GPDN01A5TLK	23	C	0.00	
FR3GPDN01A5TLK	24	G	0.16	
FR3GPDN01A5TLK	25	T	0.88	39
FR3GPDN01A5TLK	26	A	0.19	
FR3GPDN01A5TLK	27	C	0.00	
FR3GPDN01A5TLK	28	G	1.91	39,39
FR3GPDN01A5TLK	29	T	1.07	39
FR3GPDN01A5TLK	30	A	0.04	

.fna

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
TAACAATCGAGGCGAAGTCCCGTGAGAAGCTGTTTACTTCTCATGATCACACAGGCGCTG
GCTCCTCAGGCAAACAGGTACGTCTACGATAGGTTCCATGAAAAGTCCAAGTTTGGCCGA
GCTCTGGCTCCTTTTGACGCACAGTGGAACTTCCTTGTTCACGGAAATTGCA
```

.qual

```
>000007_1940_1402 length=172 uaccno=E4UQSRD01E0MP4
28 35 28 27 34 27 26 25 25 28 31 24 26 27 32 25 27 27 32 28 6  28 27 27 27 27 33 26
27 26 27 27 34 30 10 27 25 34 27 28 22 28 27 26 26 27 27 26 27 25 22 23 28 27 18 20
23 27 27 29 21 25 25 34 26 27 24 25 32 24 22 33 28 7  25 20 30 22 28 27 24 25 28 28
28 27 28 26 27 25 23 33 25 35 28 34 27 27 25 28 38 34 21 8  25 27 34 27 31 23 22 36
32 17 29 21 32 24 24 27 28 19 27 28 26 34 28 23 25 35 28 38 34 21 8  26 26 27 25 27
21 28 28 27 27 34 27 34 27 25 30 21 34 26 33 25 26 35 28 20 28 25 34 27 37 33 15 33
25 23 28 25
```

Illumina output formats

.seq.txt

```
File: s_1_0001_seq.txt
1       1       137     689     AACATAATGTGTTCACTGAGAACACATTGCACTCAA
1       1       87      649     TATTGCAACTTGTTTAATTTTTTCATGCCATTATCA
1       1       121     642     TACATGATTTGCATTTGGTAAATAGCTACTTTTTAT
1       1       6       591     C...T.......T.......................
```

.prb.txt

```
40 -40 -40 -40    40 -40 -40 -40    -40 40 -40 -40    40 -40 -40 -40
-40 -40 -40 40    40 -40 -40 -40    40 -40 -40 -40    -40 -40 -40 40
-40 -40 40 -40    -40 -40 -40 40    -40 -40 40 -40    -40 -40 -40 40
-40 -40 -40 40    -40 40 -40 -40    40 -40 -40 -40    -40 40 -40 -40
-40 -40 -40 40    -40 -40 40 -40    40 -40 -40 -40    -40 -40 40 -40
40 -40 -40 -40    40 -40 -40 -40    -40 40 -40 -40    40 -40 -40 -40
-40 40 -40 -40    40 -40 -40 -40    -40 -40 -40 40    -40 -40 -40 40
-40 -40 40 -40    -40 40 -40 -40    40 -40 -40 -40    -40 40 -40 -40
-40 -40 -40 40    -40 40 -40 -40    40 -40 -40 -40    37 -37 -40 -40
```

Illumina FASTQ

(ASCII – 64 is Illumina score)

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

Qseq

(ASCII – 64 is Phred score)

Phred quality scores

```
HWUSI-EAS521    2       1       26      0
        .GGCAGCGGGCAGGGCGAGCCAATGCGTGT(
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2       1       26      0
        .GGGAAATAGCTTTCACGCCTTTAGATAATTT(
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2       1       26      0
        .ATCTTTAACAGACCAAGACTGGGCCACAAG(
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
HWUSI-EAS521    2       1       26      0
        .GGCTGGGTATGAGTCAGGGGGCTCCAGAG
        BBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

Illumina single line format

```
>1-1-137-689     AACATAATGTTCACTGAGAACACATTGCACTCAA     U0
>1-1-87-649      TATTGCAACTTGTTTAATTTTTCATGCCATTATCA    U1
>1-1-121-642     TACATGATTTGCATTTGGTAAATAGCTACTTTTAT    U0
```

SCARF

Solexa Compact ASCII Read *Format*

```
HWI-EAS102_3:6:1:897:791:AATGTCAATCTGAGTT...TTT:40 40 40 40 40..
HWI-EAS102_3:6:1:930:291:AATGTACTTTTTCTAA...CTA:40 29 14 17 16..
HWI-EAS102_3:6:1:944:665:AATCGATCCCCTTCCC...TTC:40 34 33 40 40..
```

Illumina FastQ

```
@ILMN-GA001_3_208HWAAXX_1_1_110_812
ATACAAGCAAGTATAAGTTCGTATGCCGTCTT
+ILMN-GA001_3_208HWAAXX_1_1_110_812
hhhYhh]NYhhhhhhYIhhaZT[hYHNSPKXR
@ILMN-GA001_3_208HWAAXX_1_1_111_879
GGAGGCTGGAGTTGGGGACGTATGCGGCATAG
+ILMN-GA001_3_208HWAAXX_1_1_111_879
hSWhRNJ\hFhLdhVOhAIB@NFKD@PAB?N?
```

- ASCII value for h= 103
- Quality of Base A at the position 1 = 103- 64
- 103- 64 = 39
- Where 39 is the phred score

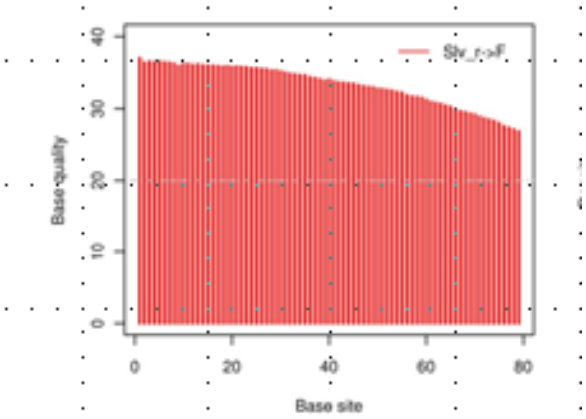
Quality Control

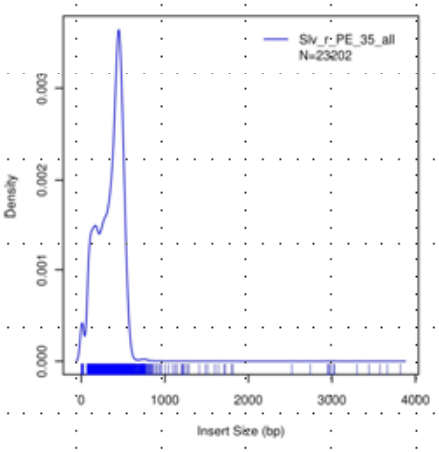
Read quality distribution

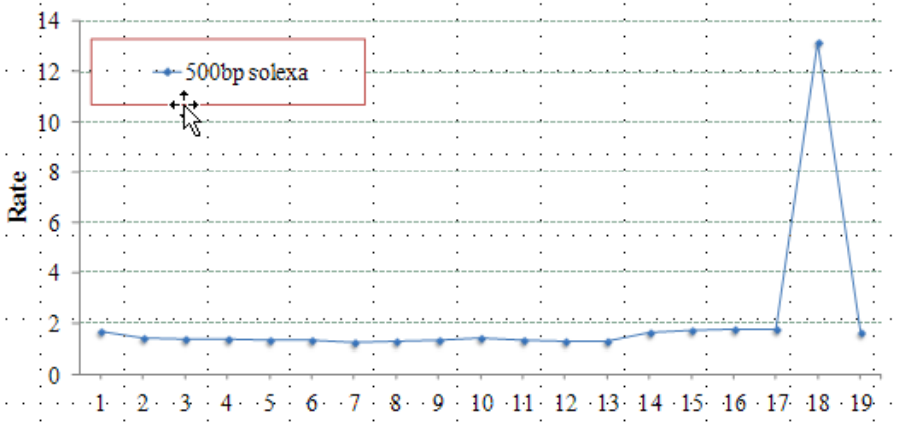
Library insert size

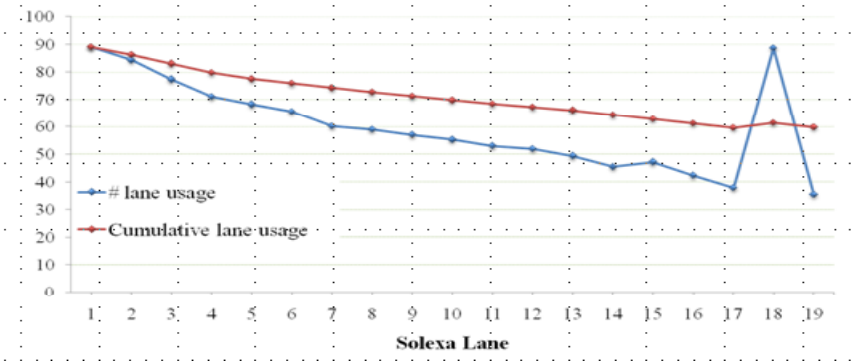
Mapping Rate

Duplication assessment









Quality Control Tools

Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17[1]	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12[1]	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ[2]	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No[3]	No	Yes	Yes	No	Yes[4]	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No[5]	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No[6]	Yes	No	Yes	No	No	Yes
Compatability with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No[7]	No	Yes	Yes	No[7]	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png	-	BLAST, NCBI nr database

NGS QC Toolkit & FastQC

- NGS QC Toolkit is for quality check and filtering of high-quality read
- This toolkit is a standalone and open source application freely available at http://www.nipgr.res.in/ngsqctoolkit.html
- Application have been implemented in Perl programming language
- QC of sequencing data generated using Roche 454 and Illumina platforms
- Additional tools to aid QC : (sequence **format converter** and **trimming tools**) and analysis (**statistics tools**)

FastQC can be used only for preliminary analysis

FastQC

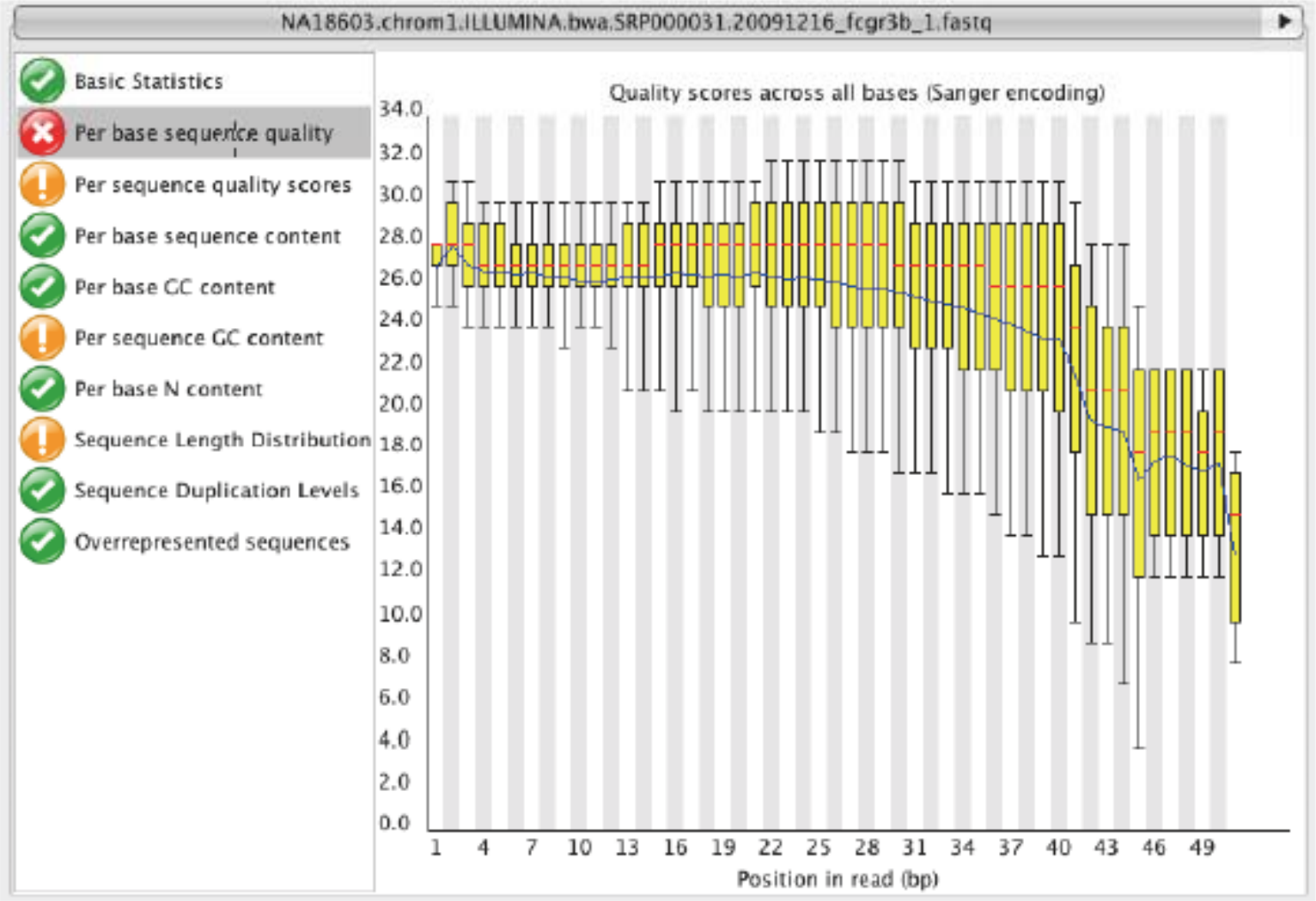


FastQC

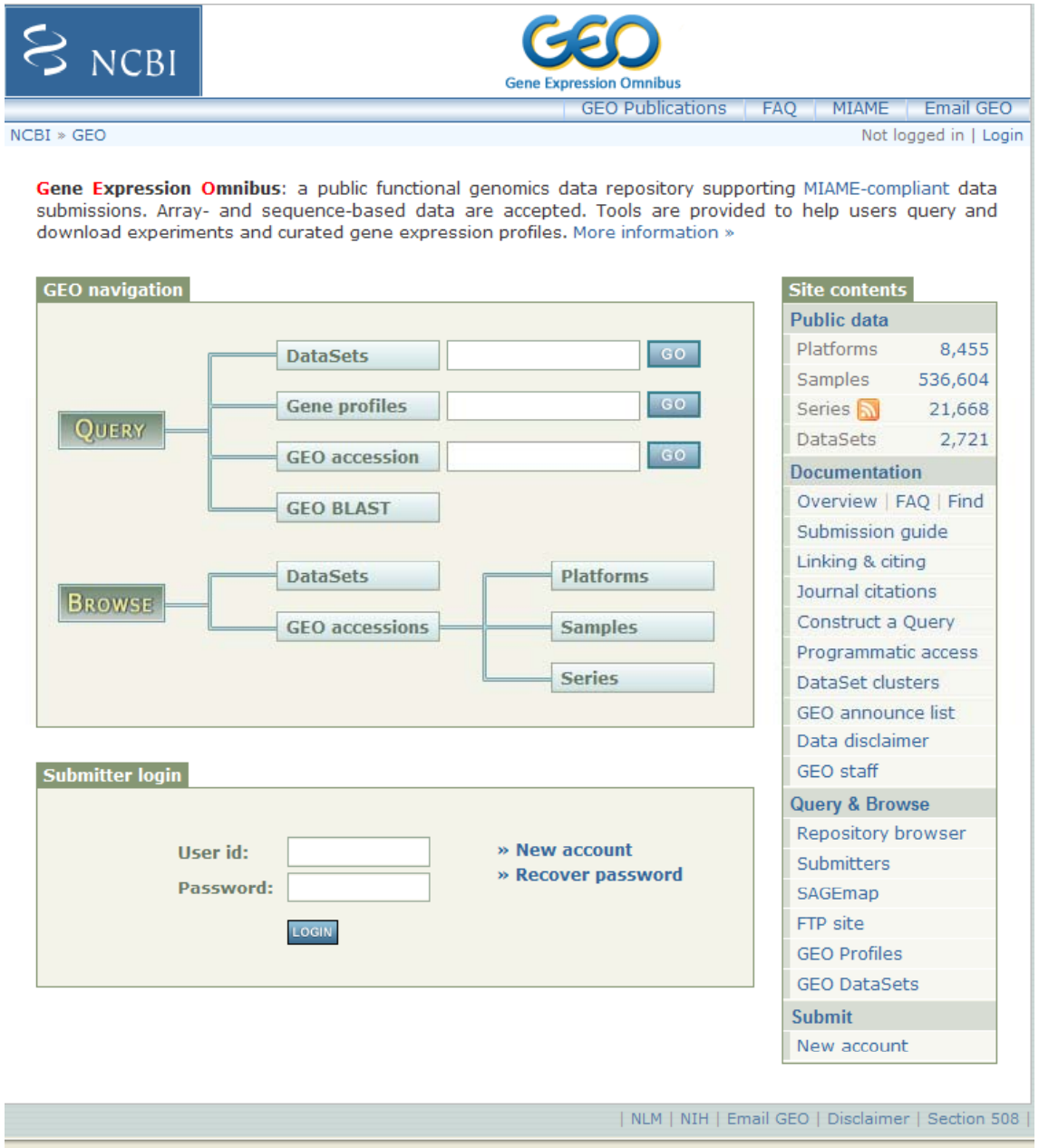
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

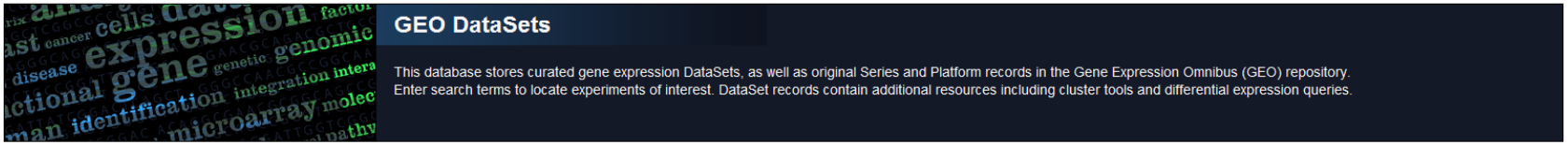
http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

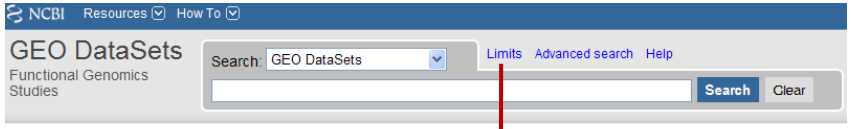

NA18603.chrom1.ILLUMINA.bwa.SRP000031.20091216_fcgr3b_1.fastq
Basic Statistics
Per base sequence quality
Per sequence quality scores
Per base sequence content
Per base GC content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Quality scores across all bases (Sanger encoding)
34.0
32.0
30.0
28.0
26.0
24.0
22.0
20.0
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49
Position in read (bp)

http://www.ncbi.nlm.nih.gov/geo/

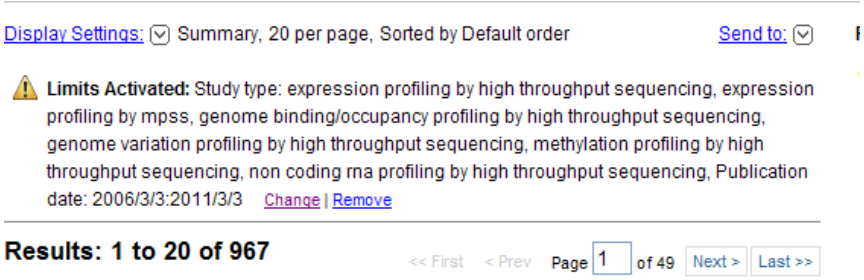


http://www.ncbi.nlm.nih.gov/gds/





expression profiling by array
expression profiling by genome tiling array
expression profiling by high throughput sequencing
expression profiling by mpss
expression profiling by rt pcr
expression profiling by sage
expression profiling by snp array
genome binding/occupancy profiling by array
genome binding/occupancy profiling by genome tiling array
genome binding/occupancy profiling by high throughput sequencing
genome binding/occupancy profiling by snp array
genome variation profiling by array
genome variation profiling by genome tiling array
genome variation profiling by high throughput sequencing
genome variation profiling by snp array
methylation profiling by array
methylation profiling by genome tiling array
methylation profiling by high throughput sequencing
methylation profiling by snp array
non coding rna profiling by array
non coding rna profiling by genome tiling array
non coding rna profiling by high throughput sequencing
other
protein profiling by mass spec
protein profiling by protein array
snp genotyping by snp array
third party reanalysis



Limits Activated: Study type: expression profiling by high throughput sequencing, expression profiling by mpss, genome binding/occupancy profiling by high throughput sequencing, genome variation profiling by high throughput sequencing, methylation profiling by high throughput sequencing, non coding rna profiling by high throughput sequencing, Publication date: 2006/3/3:2011/3/3 Change | Remove

Results: 1 to 20 of 967

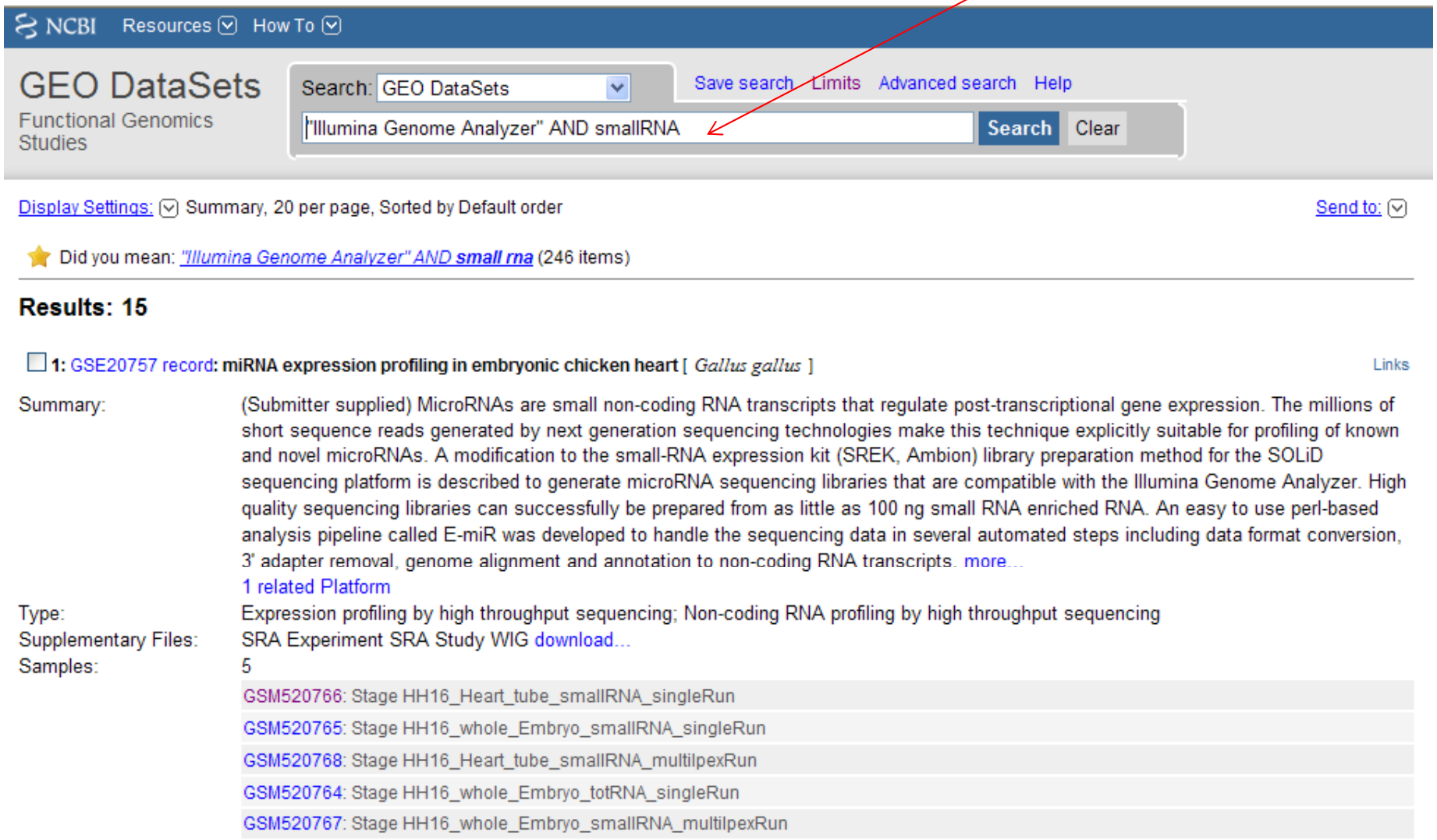
Top Organisms [Tree]

- Homo sapiens *(257)*
- Mus musculus *(193)*
- Drosophila melanogaster *(172)*
- Caenorhabditis elegans *(74)*
- Arabidopsis thaliana *(59)*
- Saccharomyces cerevisiae *(28)*
- Oryza sativa *(26)*
- Danio rerio *(12)*
- Rattus norvegicus *(11)*
- Drosophila simulans *(9)*
- Drosophila pseudoobscura *(9)*
- Zea mays *(9)*
- Gallus gallus *(8)*
- Chlamydomonas reinhardtii *(8)*
- Sus scrofa *(7)*
- Drosophila yakuba *(7)*
- Arabidopsis lyrata *(7)*
- Escherichia coli *(7)*
- Macaca mulatta *(5)*
- Xenopus (Silurana) tropicalis *(5)*

Taxonomic Groups [List]

- eukaryotes *(1086)*
  - animals *(841)*
    - chordates *(534)*
      - vertebrates *(529)*
    - arthropods *(223)*
      - flies *(205)*
    - nematodes *(77)*
  - green plants *(153)*
    - land plants *(144)*
  - fungi *(78)*
    - ascomycetes *(73)*
  - cellular slime molds *(4)*
  - apicomplexans *(3)*
  - kinetoplastids *(3)*
- viruses *(16)*
  - Herpesvirales *(6)*
  - Polyomaviridae *(3)*
  - Tombusviridae *(3)*
- other sequences *(5)*
- bacteria *(3)*

"Illumina Genome Analyzer" AND smallRNA



SRA: Sequence Read Archive

Collection of sequence data from next-generation sequencing technology for different organisms

http://www.ncbi.nlm.nih.gov/sra/ and http://www.ncbi.nlm.nih.gov/Traces/sra/

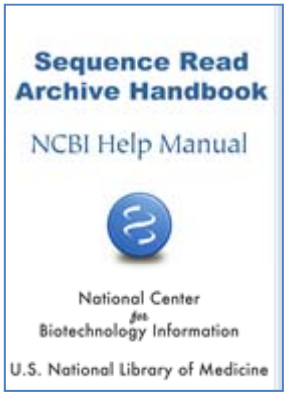
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

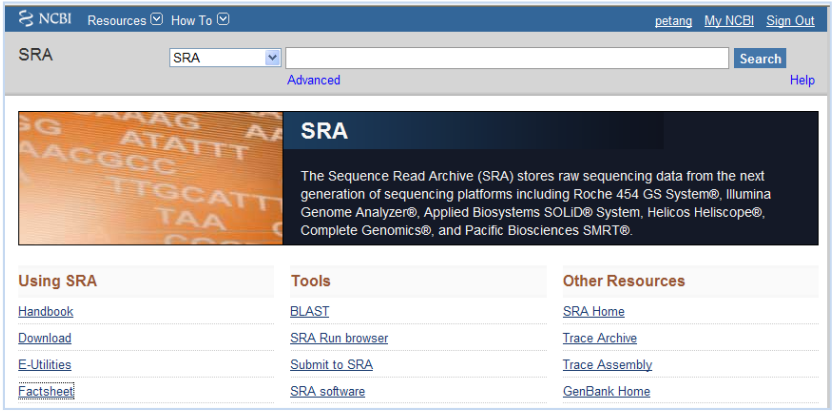
Scope and access

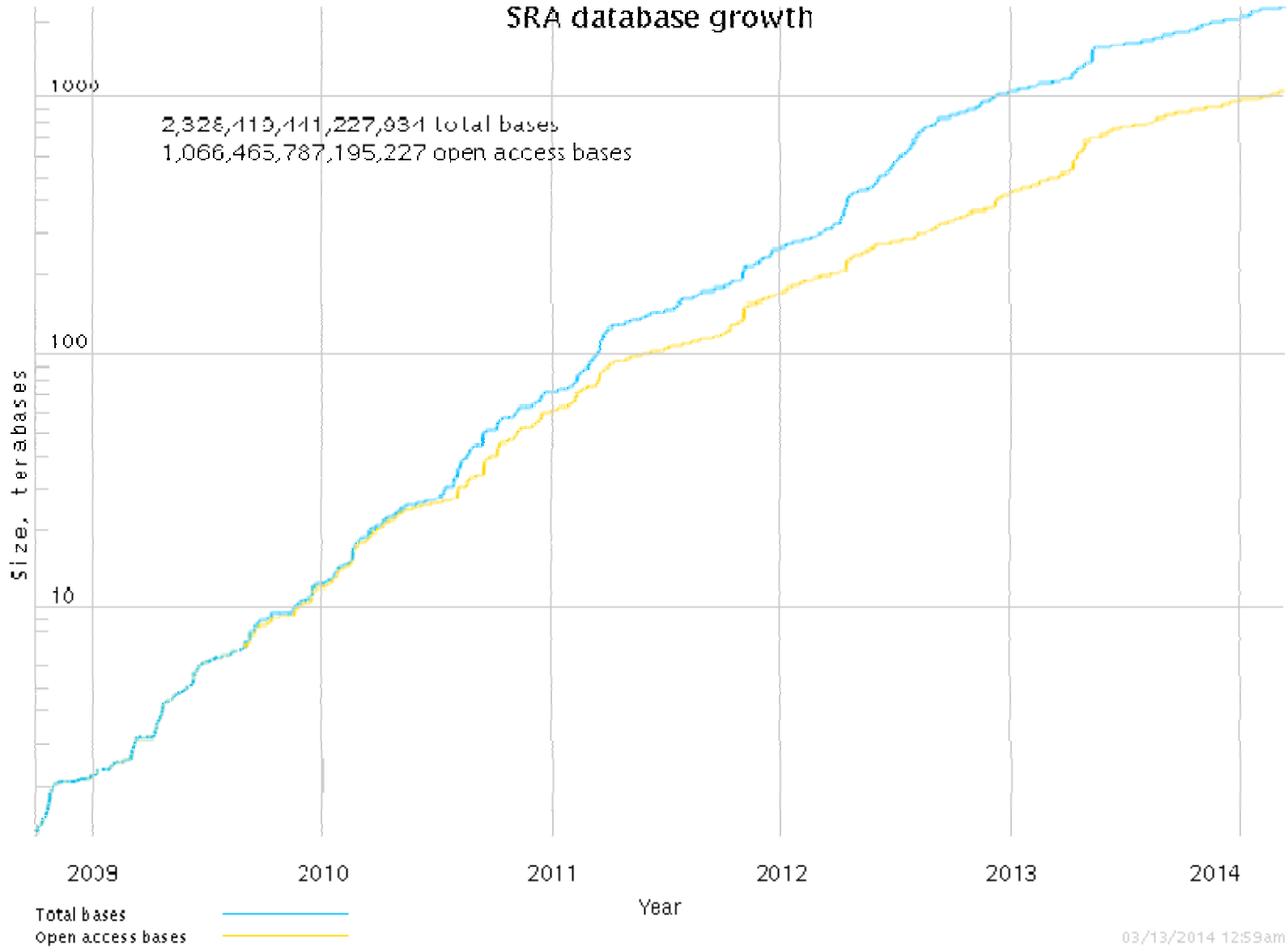
Sequence Read Archive (SRA) is the NCBI database which stores sequence data obtained from next generation sequence (NGS) technology. Through this database, the metadata for those sequences can be queried to locate the sequence dataset for subsequent download and further analysis. Specifically, SRA:

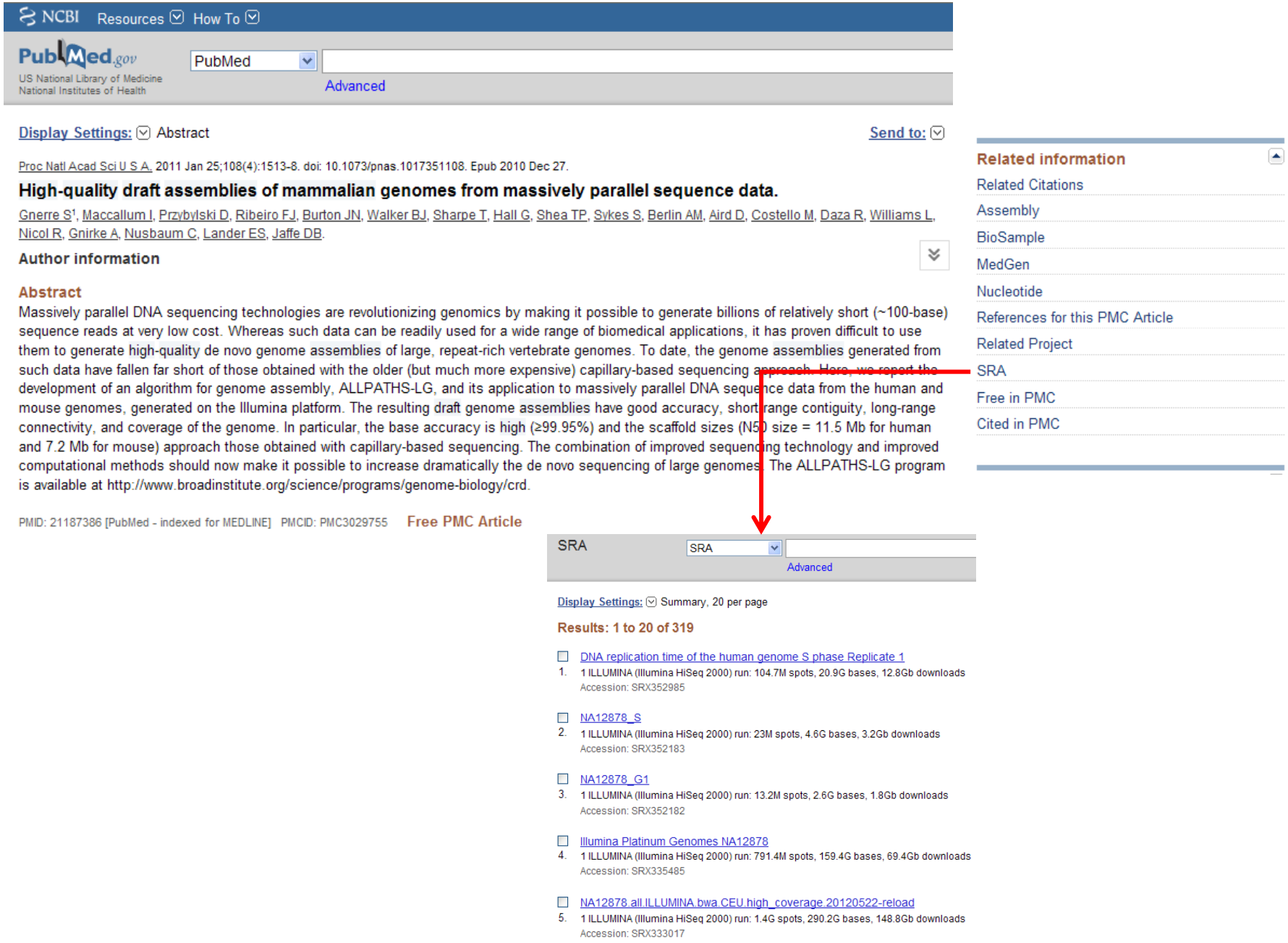
- Archives raw oversampling NGS data for various genomes from several platforms
- Shares NGS data with EMBL and DDBJ
- Serves as a starting point for "secondary analysis"
- Provides access to data from human clinical samples to authorized users who agree to the dataset's privacy and usage mandates

Metadata from SRA can be queried from Entrez SRA page and the SRA project list and sequence data can be browsed, searched and downloaded from its homepage at www.ncbi.nlm.nih.gov/sra/ and www.ncbi.nlm.nih.gov/Traces/sra/, respectively. For sequence-based searches against certain subsets of SRA reads (long reads from the Roche 454 platform) using BLAST, a link is listed under the "Specialized BLAST" section on the BLAST homepage: blast.ncbi.nlm.nih.gov/ .






SRA database growth
2,328,119,111,227,931 total bases
1,066,465,787,195,227 open access bases
1000
100
10
Size, terabases
2009
2010
2011
2012
2013
2014
Year
Total bases
Open access bases
03/13/2014 12:59am


NCBI Resources How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Advanced
Display Settings: Abstract
Send to:
Proc Natl Acad Sci U S A. 2011 Jan 25;108(4):1513-8. doi: 10.1073/pnas.1017351108. Epub 2010 Dec 27.
High-quality draft assemblies of mammalian genomes from massively parallel sequence data.
Gnerre S[1], Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB.
Author information
Abstract
Massively parallel DNA sequencing technologies are revolutionizing genomics by making it possible to generate billions of relatively short (~100-base) sequence reads at very low cost. Whereas such data can be readily used for a wide range of biomedical applications, it has proven difficult to use them to generate high-quality de novo genome assemblies of large, repeat-rich vertebrate genomes. To date, the genome assemblies generated from such data have fallen far short of those obtained with the older (but much more expensive) capillary-based sequencing approach. Here, we report the development of an algorithm for genome assembly, ALLPATHS-LG, and its application to massively parallel DNA sequence data from the human and mouse genomes, generated on the Illumina platform. The resulting draft genome assemblies have good accuracy, short-range contiguity, long-range connectivity, and coverage of the genome. In particular, the base accuracy is high (≥99.95%) and the scaffold sizes (N50 size = 11.5 Mb for human and 7.2 Mb for mouse) approach those obtained with capillary-based sequencing. The combination of improved sequencing technology and improved computational methods should now make it possible to increase dramatically the de novo sequencing of large genomes. The ALLPATHS-LG program is available at http://www.broadinstitute.org/science/programs/genome-biology/crd.
PMID: 21187386 [PubMed - indexed for MEDLINE] PMCID: PMC3029755 Free PMC Article
Related information
Related Citations
Assembly
BioSample
MedGen
Nucleotide
References for this PMC Article
Related Project
SRA
Free in PMC
Cited in PMC
SRA
SRA
Advanced
Display Settings: Summary, 20 per page
Results: 1 to 20 of 319
1. DNA replication time of the human genome S phase Replicate 1
1 ILLUMINA (Illumina HiSeq 2000) run: 104.7M spots, 20.9G bases, 12.8Gb downloads
Accession: SRX352985
2. NA12878_S
1 ILLUMINA (Illumina HiSeq 2000) run: 23M spots, 4.6G bases, 3.2Gb downloads
Accession: SRX352183
3. NA12878_G1
1 ILLUMINA (Illumina HiSeq 2000) run: 13.2M spots, 2.6G bases, 1.8Gb downloads
Accession: SRX352182
4. Illumina Platinum Genomes NA12878
1 ILLUMINA (Illumina HiSeq 2000) run: 791.4M spots, 159.4G bases, 69.4Gb downloads
Accession: SRX335485
5. NA12878.all.ILLUMINA.bwa.CEU.high_coverage.20120522-reload
1 ILLUMINA (Illumina HiSeq 2000) run: 1.4G spots, 290.2G bases, 148.8Gb downloads
Accession: SRX333017

SRX352985: **DNA replication time of the human genome S phase Replicate 1**
1 ILLUMINA (Illumina HiSeq 2000) run: 104.7M spots, 20.9G bases, 12.8Gb downloads

Accession: SRX352985
Experiment design: n/a
Submitted by: Harvard Medical School
Study summary: SRP013016 • Differential and sex-specific relationship of DNA replication timing to different forms of human mutation and variation • SRP013016 • All experiments (less...)

Abstract: Understanding the factors that affect the generation of mutations is important in medical, cancer and evolutionary genetics. DNA replication timing is known to affect the distribution of mutations across the human genome. We generated a high resolution map of the human genome's replication timing program and analyzed its relationship to different forms of human mutation and genome variation. We show that the replication program profoundly shapes the distribution of substitution mutations, with a stronger influence on transversion relative to transition mutations. In addition, we find distinct relationships between replication timing and different kinds of structural mutations: mutations mediated by non-allelic homologous recombination are enriched in the early replicating parts of the genome, while structural mutations arising from other mechanisms are enriched in late replicating regions. An early-replication bias is also observed for meiotic recombination hotspots. Both recombination and structural mutations show gender differences in their relationships to DNA replication timing, with a stronger early-replication bias observed in females. Together, our results indicate a differential and sex-specific relationship of DNA replication timing to the distribution of different forms of mutation and variation across the human genome.

Center Project:

Sample: SAMN00801888 • GM12878 female unaffected International hapmap project - ceph (plate i) utah residents with ancestry from northern and western europe sample (less...)

Organism: Homo sapiens
Attributes:
cell line: GM12878
dna id: NA12878
sex: Female
cell type: B-Lymphocyte
family id: 1463
family role: mother
race: Caucasian
ethnicity: UTAH/MORMON
collection: NIGMS Human Genetic Cell Repository
tag: Cytochrome p450, subfamily iic, polypeptide 19
tag: Cyp2c19
tag: International hapmap project - ceph utah residents with ancestry from northern and western europe
tag: International hapmap project - ceph (plate i) utah residents with ancestry from northern and western europe
tag: Ceph/utah pedigree 1463

Library: Replicate 1 (less...)
Strategy: WGS
Source: GENOMIC
Selection: size fractionation
Layout: PAIRED

Platform: Illumina (less...)
Instrument model: Illumina HiSeq 2000

Spot descriptor:



Total: 1 run, 104.7M spots, 20.9G bases, 12.8Gb

#	Run	# of Spots	# of Bases	Size
1.	SRR496477	104,650,130	20.9G	12.8Gb

ID: 501159

http://seqanswers.com/

