National Cancer Institute

Introduction to Data Processing and Analysis for Illumina Platforms at CCR-SF

Bioinformatics Training and Education Program
Office of Science & Technology Partnerships
(OSTP) CCR, NCI

Yongmei Zhao
Bioinformatics Analyst III
Advanced Biomedical Computing Center, SAIC-Frederick, Inc.
Frederick National Laboratory for Cancer Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

SAIC Frederick

Goals

- Overview of Illumina sequencing technology as implemented at the CCR-SF
- Overview of CCR-SF sample QC and data analysis workflow steps
- Provide guidance for file navigation for files/data provided by CCR-SF
- Review subsequent bioinformatics options and project submission process

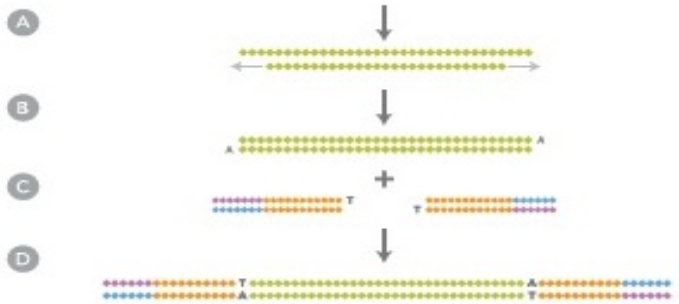
History of SAIC-F/CCR Sequencing Facility

- Opened in March 2009 and commenced service in June 2009 (three GA II instruments)
- Currently operate (2) GA IIx, (3) HiSeq 2000, and (1) PacBio RS
- A MiSeq is currently in testing

Overview Illumina Sequencing Technology

1 LIBRARY PREPARATION

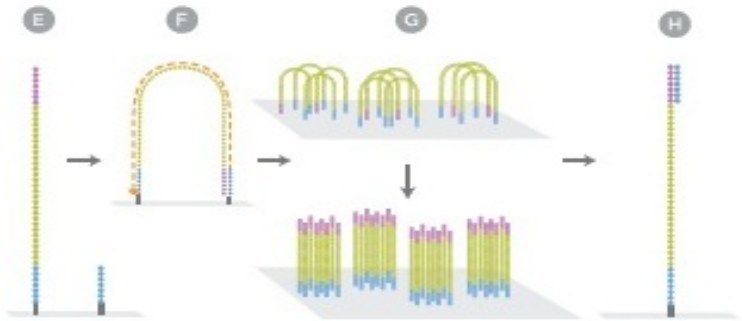
6 hours
3 hours hands-on time



- A Fragment DNA
- B Repair ends
  Add A overhang
- C Ligate adapters
- D Select ligated DNA

2 CLUSTER GENERATION

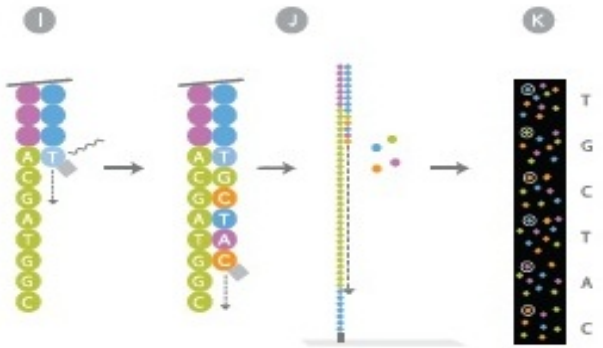
4 hours
5 minutes hands-on time
1–96 samples



- E Attach DNA to
  flow cell
- F Perform bridge
  amplification
- G Generate clusters
- H Anneal sequencing
  primer

3 SEQUENCING

1–3 days single-read run
3–7 days paired-end run
30 minutes hands-on time
8 lanes, up to 96 samples
per flow cell (run)



- I Extend first base,
  read, and deblock
- J Repeat step above
  to extend strand
- K Generate base calls

CCR-SF Supported Sequencing Applications

- **DNA/RNA-Protein Interactions**
  - Immunoprecipitation of DNA or RNA-binding proteins using Chip-Seq, RIP-Seq
  - Analysis of the binding sites of transcription factors, replication and transcriptional machinery. Structural proteins such as histones, and the impact of protein modifications on genome occupancy.
- **Transcript Profiling and Discovery**
  - mRNA-seq, tag profiling, and small RNA analysis.
  - Measurement of alternative isoforms, discovery of novel structures.
- **Genome-wide Methylation Study**
  - Detection of variations in methylations signature at single-base resolution such as using whole-genome bisulfite-seq.
- **Targeted Exome or Custom Enrichment Sequencing**
  - Targeted resequencing focuses on a subset of the genome

CCR-SF Standard Data QC Metrics

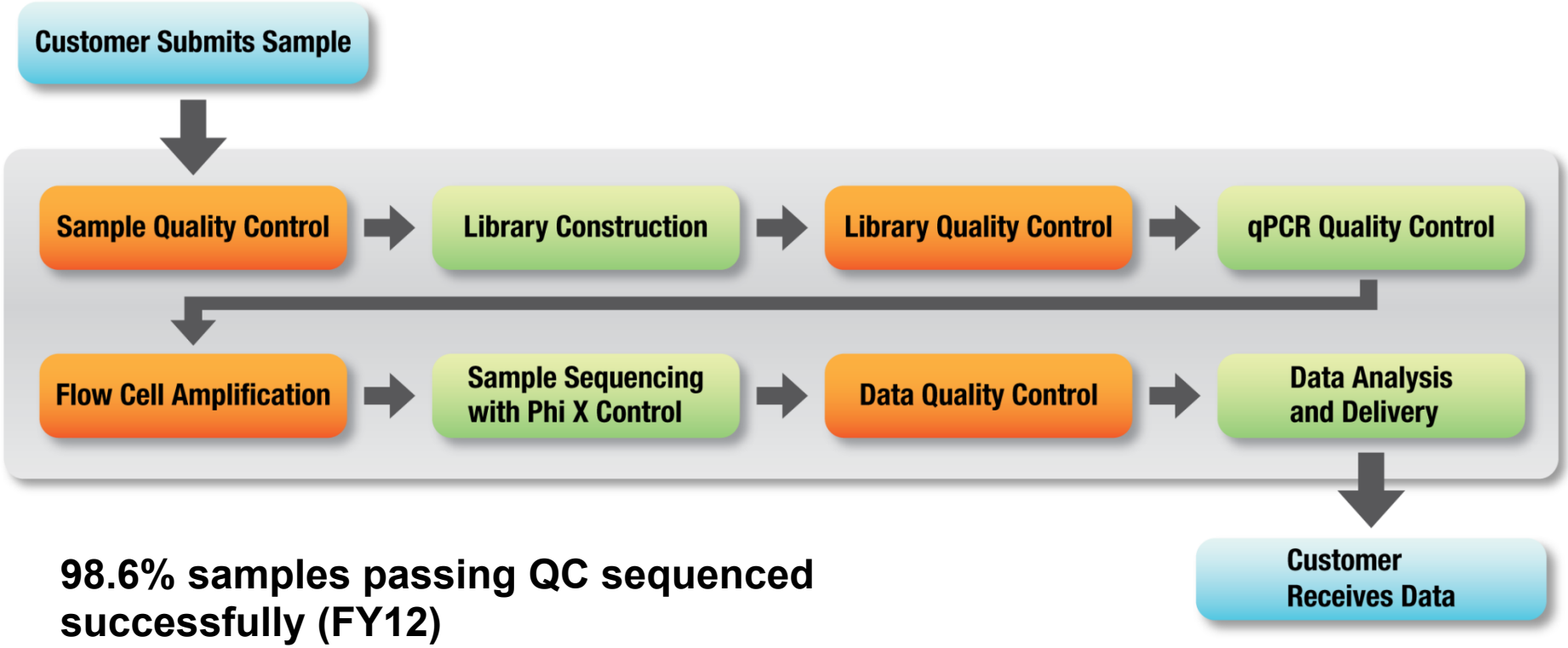
Platform	GAIIx- SR 36bps				
Application Type	Total PF Reads(Million)	Total Yield (Gb)	%>=Q30	%Alignment (PF)	%Error Rate
ChipSeq	25 - 35	0.8 -1.5	>= 85%	>= 60%	<= 0.5%
PCR- product	25 - 35	0.8 -1.5	>= 85%	>= 60%	<= 0.5%
Other	20 - 35	0.7 - 1.5	>= 85%	>= 50%	<= 1%

Platform	HiSeq2000 - PE 2x101 bps				
Application Type	Total PF Reads (Million)	Total Yield (Gb)	%>=Q30 (PF)	%Alignment (PF)	%Error Rate
mRNA	250 - 340	25 - 34	>= 75%	>= 60%	< 2%
Exome	280 - 360	28 - 36	>= 80%	>= 75%	< 2%
Other	250 - 360	25 - 36	>= 75%	>= 60%	< 2%

Note: Statistics collected is based on TruSeq V3 chemistry and software used in Q2 of 2012 by using Illumina sample prep protocols. Only pass filter (PF) reads are used. mRNA Library performance may be difference if use different library prep kits.

SF Quality Control Process

At CCR–SF, producing high-quality data is our highest priority.
To ensure this, we have quality control checks in place throughout the entire production pipeline.



98.6% samples passing QC sequenced successfully (FY12)

Illumina Data QC and Analysis Workflow

Run Time Data Processing and QC

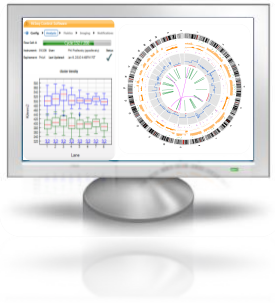
CCR-SF LIMS Capture Project and Sample Info → Generate Sample Sheet → Instrument Image Analysis and Basecalling → Illumina RTA or SAV Software → Run Time QC Analysis via SAV and RTA Reports → Illumina or Third Party Software → Optional QC Analysis with Selected Subset Data

Current Data Storage at ATRF and Computing on ABCC or ATRF Clusters at SAIC-Frederick

Post Sequencing Run Data Analysis

Demultiplex and Data Conversion → Illumina CASAVA Pipeline → Application Specific Mapping and Data Conversion → Inhouse scripts and ABCC-Hosted Software → Data Quality Assessment and Quality Matrix Generation → Inhouse scripts and ABCC Hosted Software → Report Generation, Data Archive and Data Delivery

Data Delivery to Customers via LIMS or FTP

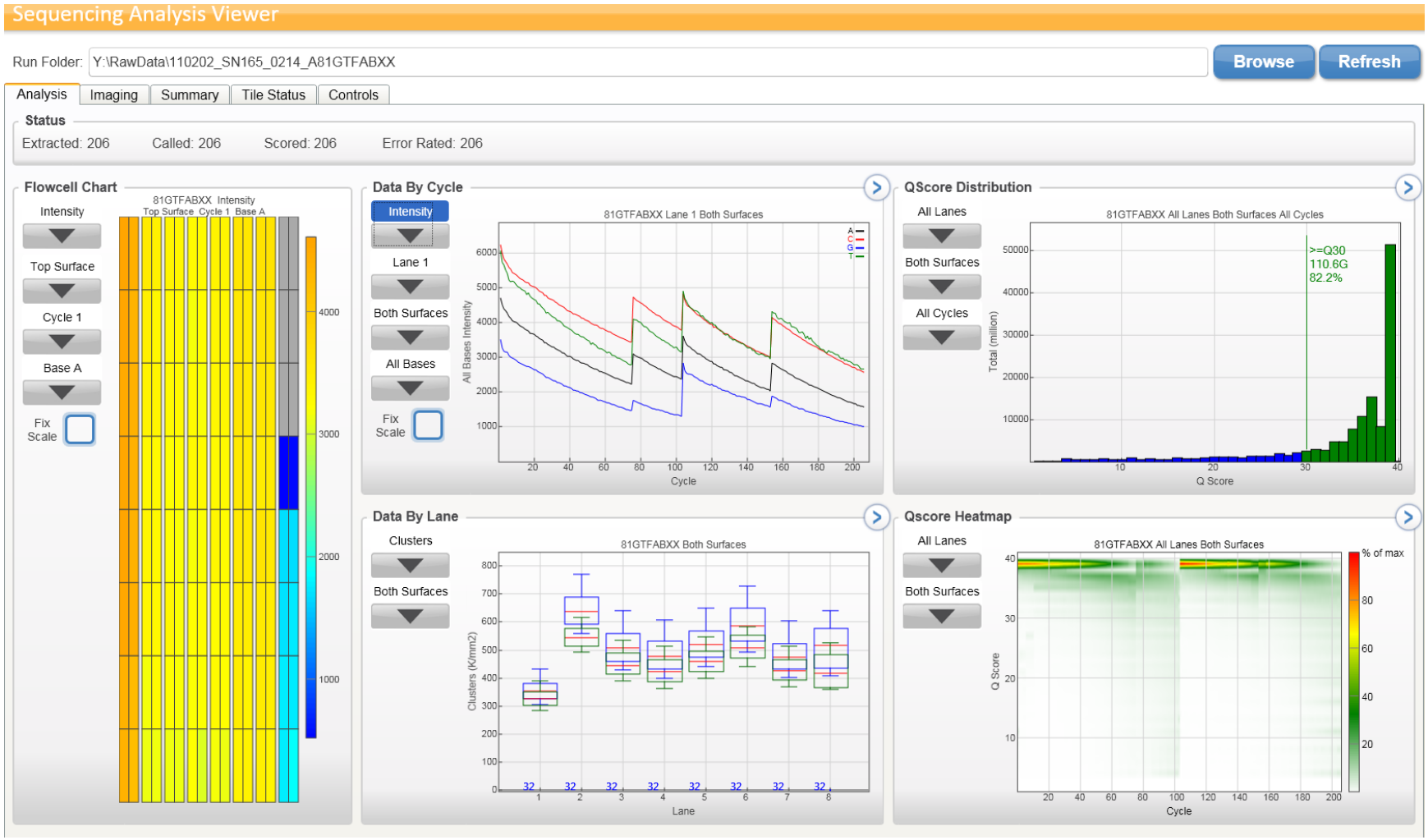


Downstream analysis at CCR-IFX Core or PI labs, and Biological Interpretation

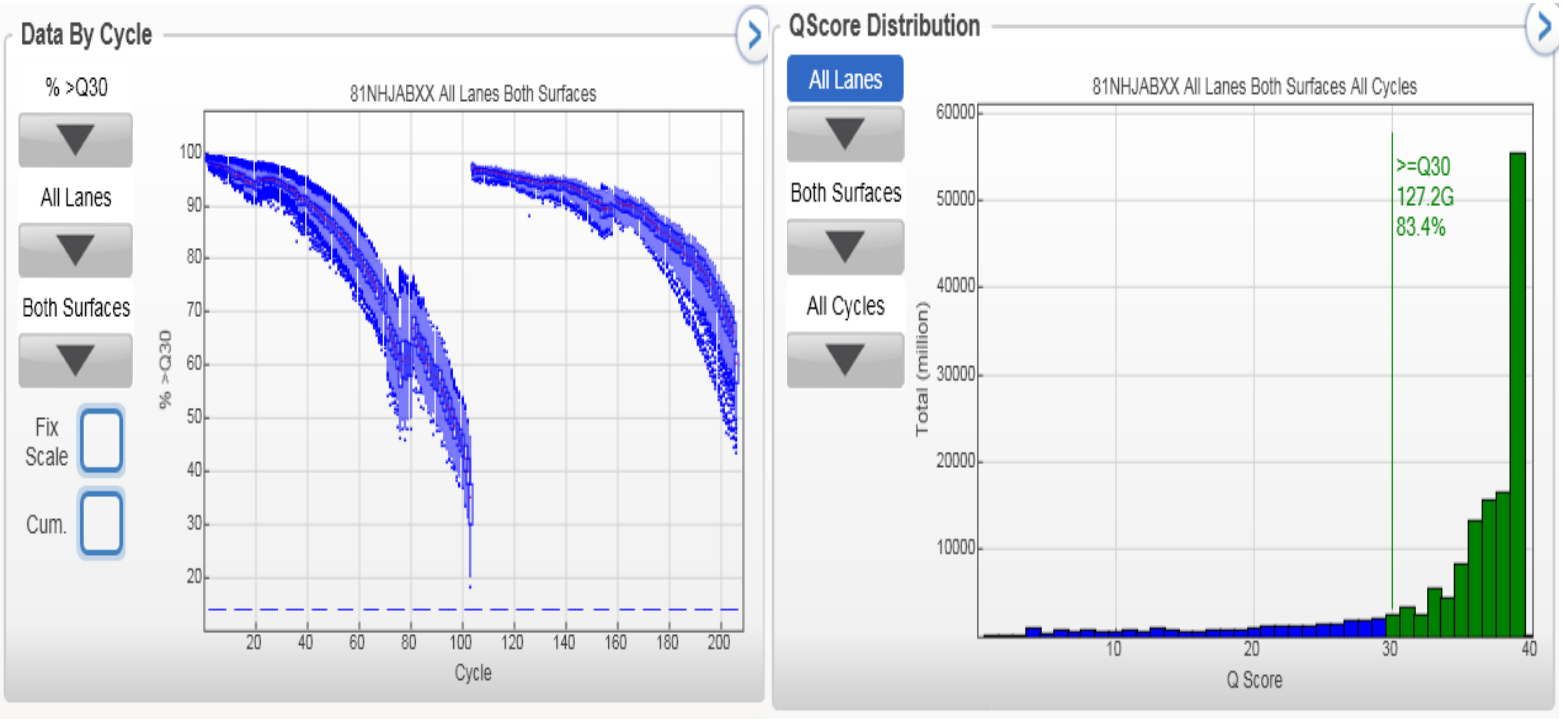
Instrument Run Time QC Parameters to check

- Cluster Intensities: Box plots
- Chemistry intensities: Cycle by cycle intensities, IVC plots
- Phasing and Prephasing
- Quality scores: cycle by cycle qscore as well as cumulative distribution of %>Q30 among reads
- Focus matrix
- Clusters raw image files

HiSeq HCS or SAV



Qscore Assessment



Illumina Quality Score

- The quality scoring scheme is the Phred scoring scheme, encoded as an ASCII character by adding different offset(64 or 33) to the Phred score.
- A Phred score of a base is: Qphred =-10 $\log_{10}(P)$ where p is the estimated probability of a base being wrong.

Phred Quality Score	Prob of Incorrect Basecall	Basecall Accuracy	QScore
20	1 in 100	99%	Q20
30	1 in 1000	99.9%	Q30
40	1 in 10,000	99.99%	Q40

Overview of Illumina Data Analysis Workflow

- Demultiplex pooled samples
- Quality assessment with third party tools
- Adapter clips and low quality bases trimming
- Alignment with ELAND or other mapping software tools
- BAM file statistics collection – Picard/SAMtools/ Bamtools
- Post alignment application specific tertiary analysis

Demultiplex and BCL Conversion

- **Demultiplex and Convert Base Calls to Compressed FASTQ Files with CASAVA configureBclToFastq.pl Script.**

  - **Demultiplex command:**

configureBclToFastq.pl --sample-sheet sample_sheet.csv --output-dir ./Unaligned --input-dir path_to_Instrument_run_dir/Data/Intensities/BaseCalls --ignore-missing-bcl –ignore-missing-stats

--use-bases-mask Y100n*,I*,I*,Y100n* #for demultiplex dual index run

  - **Sample_sheet.csv**

FCID,Lane,SampleID,SampleRef,Index Seq,Description,Control,Recipe,Operator,SampleProject
D1JF9ACXX,1,Phix,Phix,,Phix control,Y,PE_cBot_Full_Process,Lena,Phix_control
D1JF9ACXX,2,MT19,Human,GCCAAT,exome capture,N,PE_cBot_Full_Process,Lena,project_101898
D1JF9ACXX,2,MN19,Human,CAGATC,exome capture,N,PE_cBot_Full_Process,Lena, project_101898

D1JF9ACXX,7,MT20,Human,ACTTGA,exome capture,N,PE_cBot_Full_Process,Lena, project_101898
D1JF9ACXX,8,MN20,Human,CAGATC,exome capture,N,PE_cBot_Full_Process,Lena,project_1018984

Quality Assessment

- **Quality assessment with third party tools:**
  - **FASTQC, RSeQC, PRINSEQ, NGSQC**

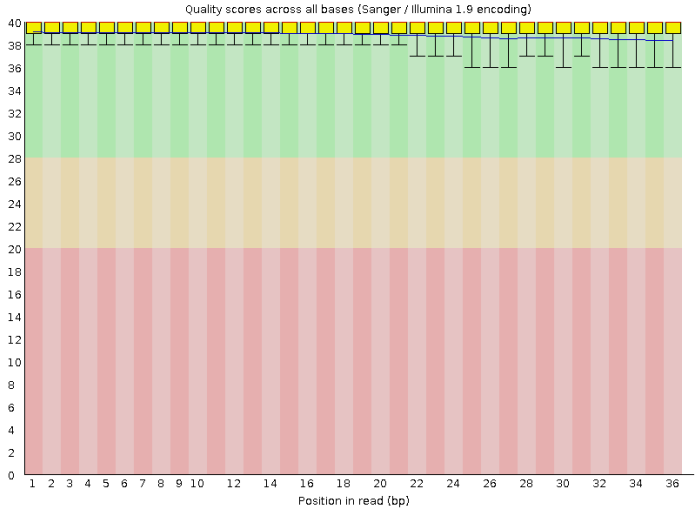
fastqc command:

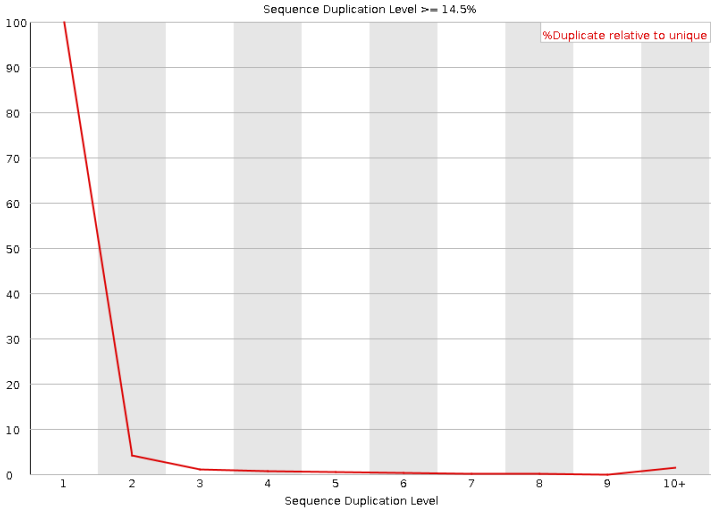
fastqc -o output_dir -f fastq -c contaminant_file seqfile1.fastq.gz .. seqfileN.fastq.gz

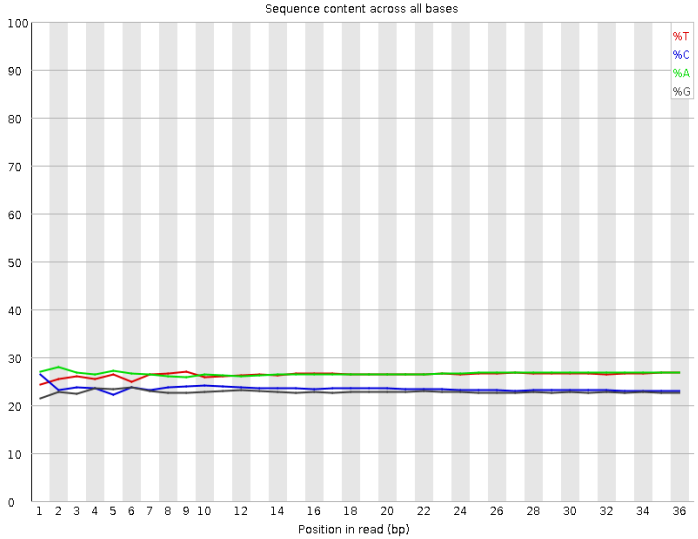
Fastqc QC metrics

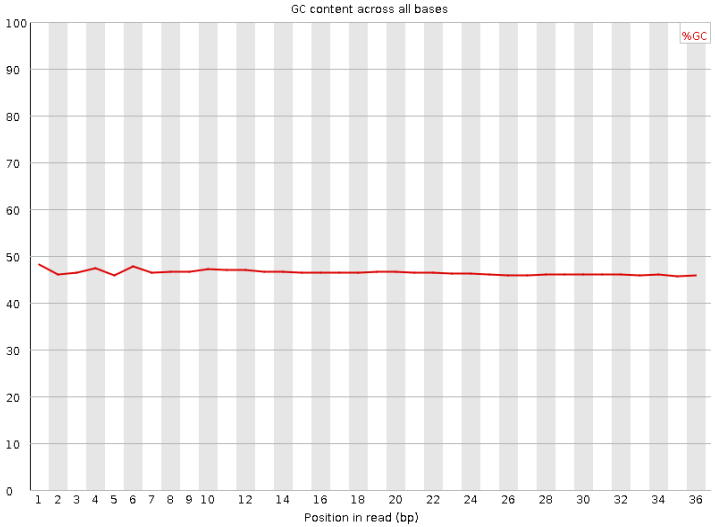
- ***Per base and per sequence quality scores:***
- ***Per base sequence content***
- ***Per base and per sequence GC content:.***
- ***Duplicate sequences***
- ***Overrepresented sequences.***
- ***Overrepresented Kmers:***

FASTQC Results



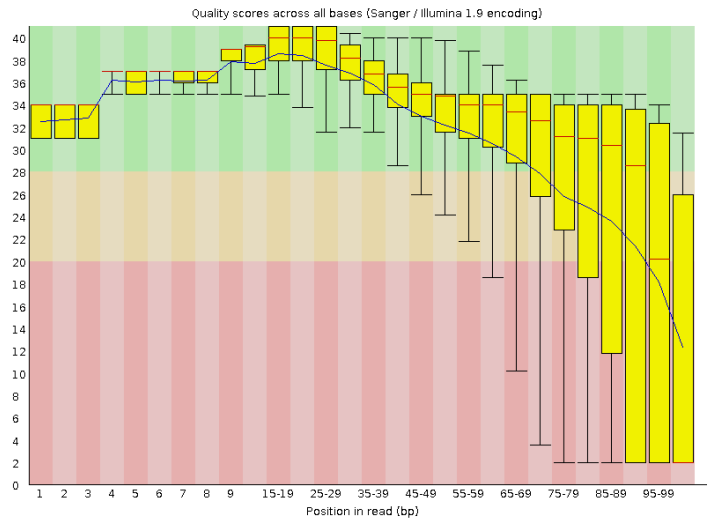




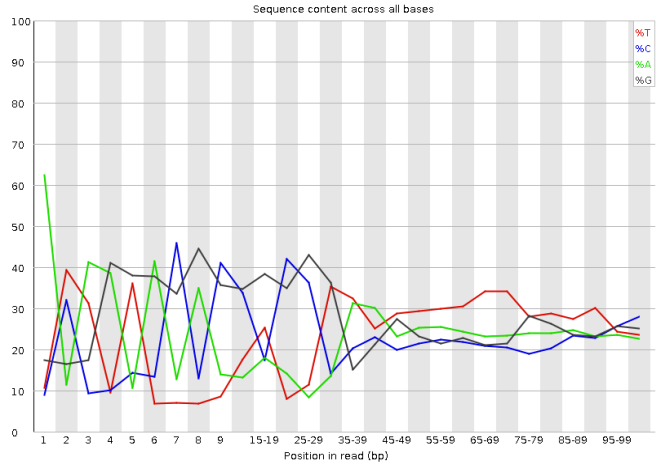


FASTQC Results

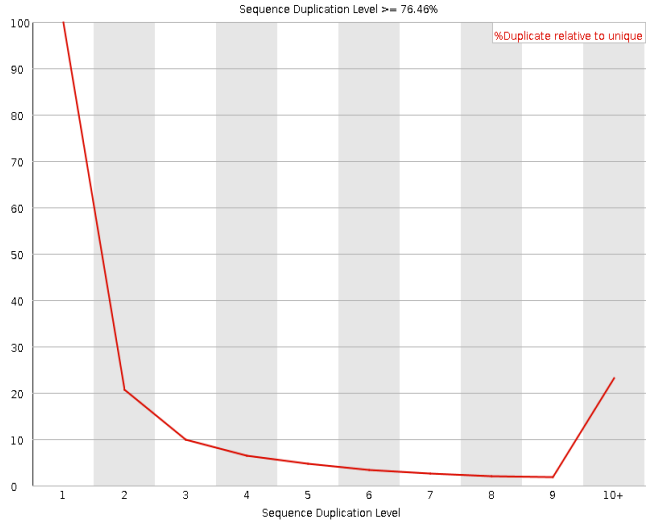
Per base sequence quality



Per base sequence content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTAATACGACTCACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTA	207987	5.199675	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGTGGG	187358	4.68395	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGT	124297	3.107425	No Hit
TGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGC	44584	1.1146	No Hit
AGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGTGGGTC	30608	0.7652	No Hit
TAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGTGGGT	20695	0.517375	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACTCCAGCTTG	18015	0.45037499999999997	No Hit
GGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGTGGGTCT	16406	0.41015	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACTTTGGGAGG	15965	0.399125	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACAGGTTTCTGGAG	10697	0.267425	No Hit
GGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATAAATCGTGGGTCTT	9852	0.2463	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACAGGTTTCTG	9036	0.22590000000000002	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATTTAACCC	6193	0.154825	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACCTGATATGG	6156	0.1539	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACTTTGGGAGGCTG	5980	0.1495	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACTCCAGCCTG	5889	0.147225	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACATTTAACCCAGT	5751	0.14377500000000001	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACAGTGGCTCA	5333	0.133325	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACTTTGGGAGGCCG	4819	0.12047500000000001	No Hit
ACTATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACCCTCTGATA	4700	0.11750000000000001	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACCAAAGTTCCAGA	4174	0.10435	No Hit
ATAGGGCACGCGTGGTCGACGGCCCGGGCTGGTTATACCAGAATCCCAGA	4020	0.1005	No Hit

Adapters and Low Quality Trimming

- **Adapter Clips and Low Quality Bases Trimming**
  - **FASTX_TOOLKIT(fastx_clipper, fastx_trimmer), fastq-mcf, cutadapt, or Trimmomatic**

  - **Fastx_clipper for adapter trimming:**

Fastx_clipper –a adapter_sequence -l 15 -l seq_R1.fastq –o seq_trimmed_fastq >run_fastx_clipper.log 2>&1 &

  - **Fastq-mcf adapters and low quality end trimming:**

fastq-mcf –o seq_R1.trimmed.fastq –o seq_R2.trimmed.fastq -l 30 -q 20 -x 10 -u -P 33 illumina_adapters_list.fa seq_R1_raw.fastq seq2_R2_raw.fastq >run_fastq-mcf.log 2>&1 &

    - **Illumina Adapters files can include:**
      - Illumina single-end/pair-end adapters
      - Illumina Single-end/pair-end PCR primers
      - Illumina TruSeq universal adapters and index adapters
      - Illumina small RNA adapter and PCR primers.

  - **Cutadapt for remove adapter at both 5' and 3' ends:**

cutadapt –g adapter_seq1 –g adapter_seq2 seqs.fastq --overlap=18 --minimum-length=20 –o seqs_adpt.trimmed.fastq --rest-file=filtered_seqs.trimmed_rest --too-short-output=filtered_seqs.fastq.2short --untrimmed-output=filtered_seqs.fastq.untrimmed >cut_adapt_run.log 2>&1 &

Mapping Software Tools

Tools for mapping high-throughput sequencing data

Fonseca et al. Bioinformatics Oct, 2012

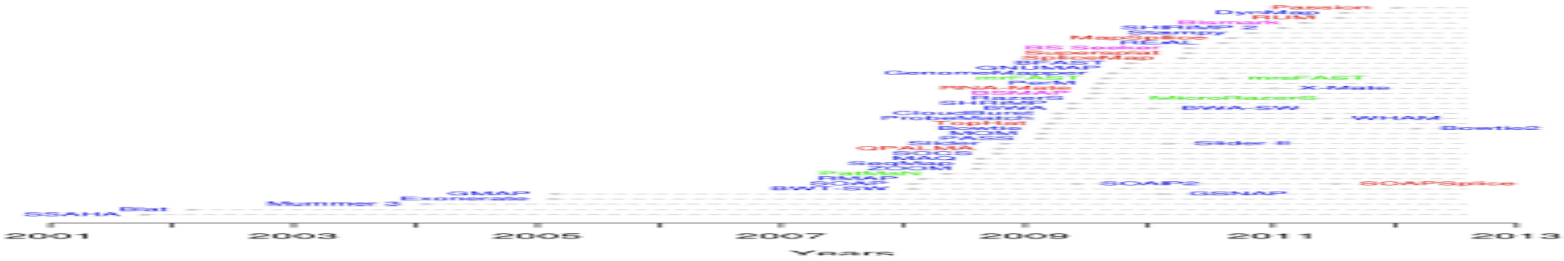


Fig. 1. Mappers time line (since 2001). DNA mappers are plotted in blue, RNA mappers in red, miRNA mappers in green, and bisulfite mappers in purple. Gray dotted lines connect related mappers (extensions or new versions). The time line only includes mappers with peer-reviewed publications and the date corresponds to the earliest date of publication (e.g., advanced publication date as opposed to the date of publication).

mrsBAST	25	200	Y	0	N	A	G	N	N	Y	N	miRNA
Mummer 3	10	*	Y	Y	Y	A,B	G	N	N	N	N	DNA
Novoalign	30	300	8	2	N	A, B, R, U, S	G	SM/DM/Cloud	Y	Y	Lib	DNA
PASS	23	1K	Y	Y	Y	A,B	G	SM	Y	Y	De novo	DNA
Passion	-	-	Y	Y	Y	U	-	SM	Y	Y	De novo	RNA
PatMaN	1	*	Y	Y	N	A	G	N	N	N	N	miRNA
PerM	20	128	9	0	Y	A,U	G	DM	Y	Y	N	DNA
ProbeMatch	36	50	3	Y	N	A,B		N	N	N	N	DNA
QPALMA	-	-	Y	Y	Y	B	L	N	Y	N	Lib and de novo	RNA
RazerS	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N	DNA
REAL	4	*	Score	N	N	B, U	G	SM	Y	N	N	DNA
RMAP	11	10K	Y	0	N	B,S		N	Y	Y	N	DNA
RNA-Mate	-	-	Y	0	N	S	-	DM	Y	N	Lib	RNA
RUM	-	-	Y	Y	Y	B	-	SM	N	Y	De novo	RNA
SeqMap	15	500	5	3	N	A		SM	N	N	N	DNA
SHRiMP	14	1K	Score	Score	Y	B,S	G	SM	N	Y	N	DNA
SHRiMP 2	30	1K	Y	Score	N	B,U,S	G	SM	Y	Y	N	DNA
Slider		62	3	0	N	B,S		N	Y	Y	N	DNA
Slider II		93	Y		N	B,S		N	N	Y	N	DNA
Smalt	4	2048M	Score	Score	N	A,B,R,U,S	L	SM	Y	Y	N	DNA

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

SF Currently Supported Mapping Software

- CASAVA ELANDv2e
  - Multiseed and gapped alignment support both DNA and RNA data.
- Bowtie/Bowtie2
  - Ultrafast, memory-efficient short read aligner using Burrows-Wheeler indexing.
  - Bowtie uses ungapped alignment and Bowtie2 supports gapped alignment perform better for reads longer than 50bp.
- TopHat
  - Fast splice junction mapper for RNA-seq reads. Uses either Bowtie or Bowtie2 as aligner.
- BWA
  - Burrows-Wheeler Aligner uses Burrows-Wheeler Transform plus auxiliary data structures which enables fast exact matching. Support both short read and long-read mapping (BWA-SW).
- Bismark
  - Map bisulfite treated sequencing reads to a genome of interest and perform methylation calls in a single step. Use bowtie/bowtie2 as underline aligner

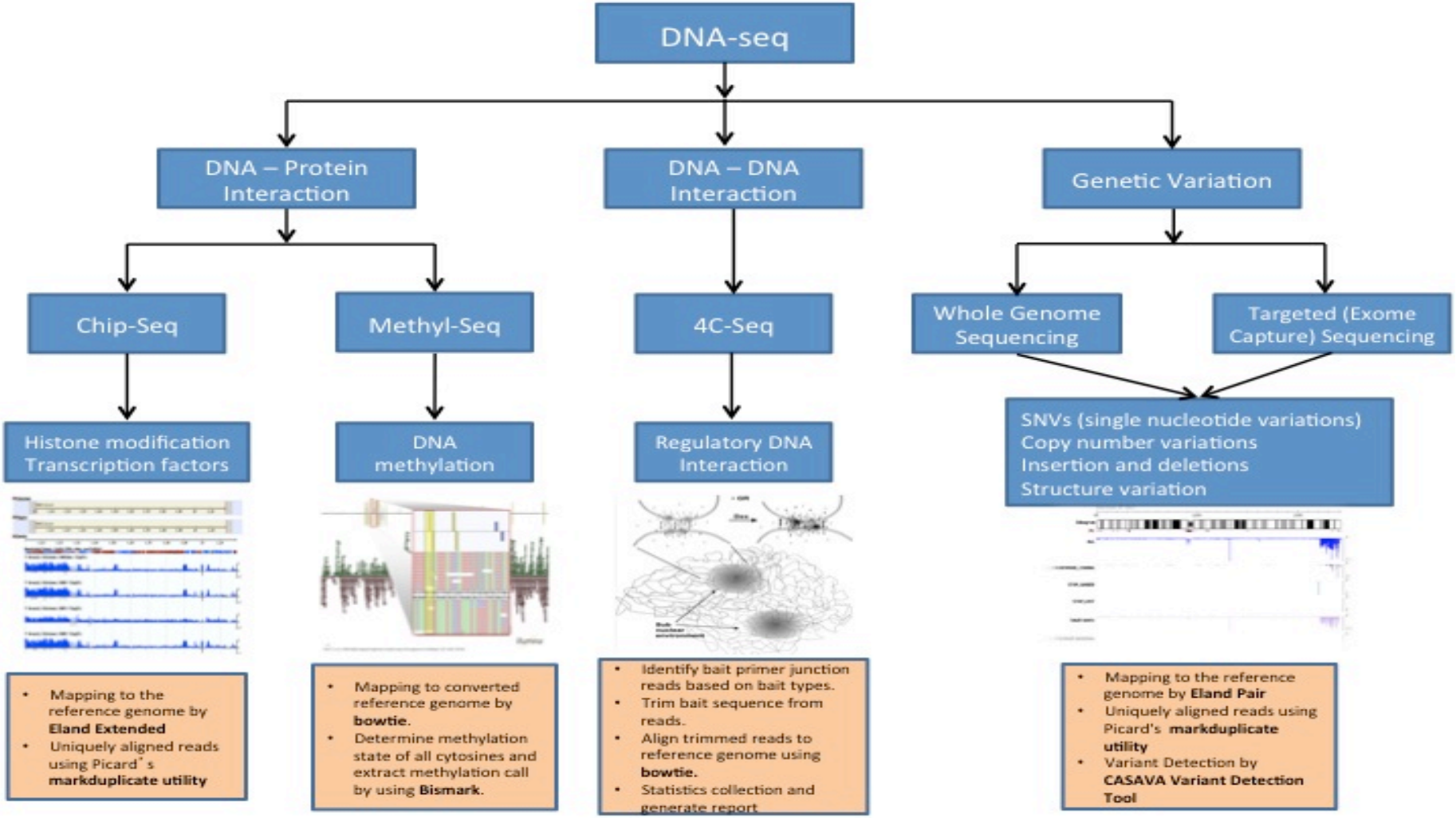
SF Currently Supported Mapping Software

Mapper	Data	Mismatches	Indels	Gaps	Align. Reported	Alignment	Parallel	Splicing
ELANDv2e	DNA /RNA	2 in seed	Score	Y	B	G	Y	Y
Bowtie	DNA	Score	Score	N	A,B,R,S	G L	SM	N
Bowtie2	DNA	Score	Score	Y	A,B,R,S	GL	SM	N
BWA	DNA	Y	8	Y	R,S	G	SM	N
TopHat	RNA	5	2	N/Y	B,S		SM	De novo
Bismark	Bisulfite	Score	Score	N	U		SM	N

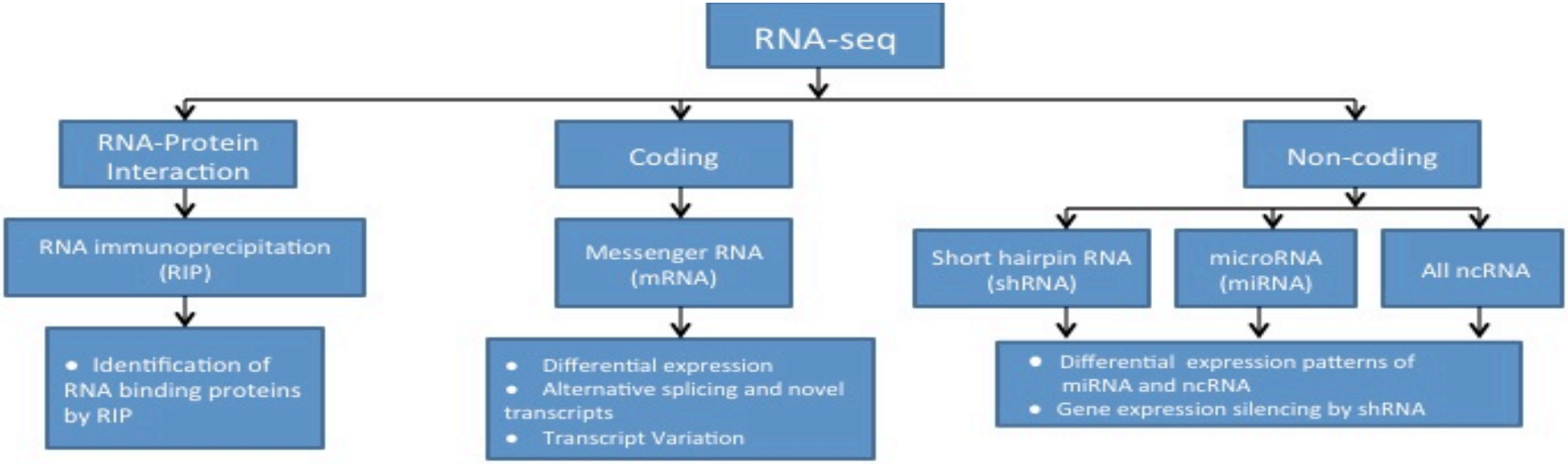
Alignments reported: A-all, B-best; R-random; U-unique alignments only (no multimaps); S-user defined number of matches.
Alignment: G-(semi-)global (a.k.a. end-to-end); L-Local.
Parallelism: SM-shared-memory; DM-distributed memory.
Y-Yes; N-No

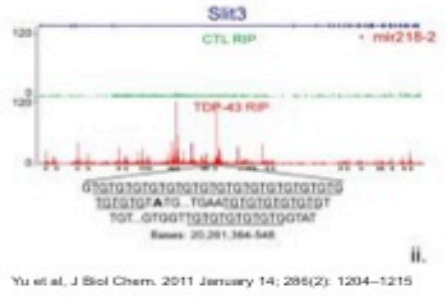
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Application Specific Analysis Pipeline

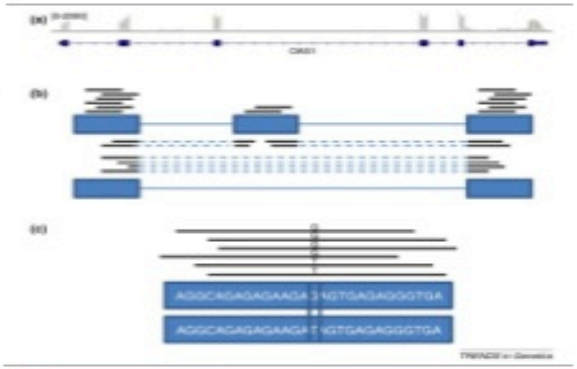


Application Specific Analysis Pipeline

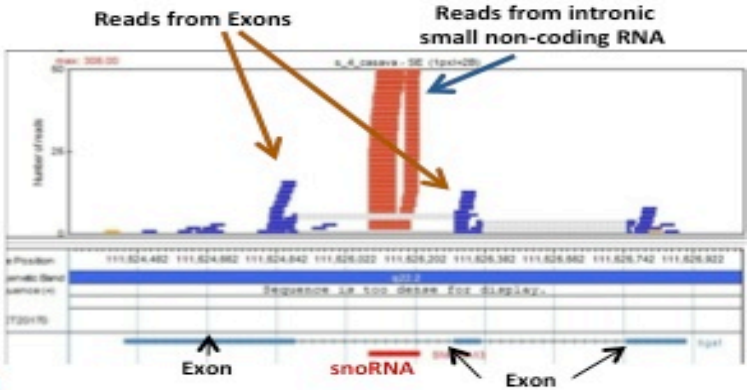




• Remove sequence adapters using fastx toolkit
•RIP sequence alignment using TopHat



• Alignment to the reference genome and known splice junctions using CASAVA eland pipeline
• Picard collectRNASeqMetrics to identify alignment to different functional classes loci ie. Coding, intronic, UTR, intergenic



• Remove sequence adapters using fastx toolkit
•Alignment to the reference genome and specific databases using CASAVA eland pipeline for shRNA and miRNA

CASAVA Output Files

- Pre-CASAVA1.8

- Qseq.txt: Raw fastq file, contains all reads. Each read has one line (quality encoding is phred64)

```
NCI-GA4 0004    5       93      1050    19307   0       1
AACATGGCCTAAAGGCAGATTTTGAAGCGGTTGTAGGGGGCAAGAGCCCTATTCGTCTTTAAATTACTTGATATCA    KWX\
\RaKK^SINSHXXZHP``_BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB    0
NCI-GA4 0004    5       93      1049    16589   0       1
CGAATAGCCTTTGTGCATTGTTTTGGACTGAGATCTGCTTCATTAGCAAATTTCAACTATATGATACAGTATTTCT    effceffffedcffeefffcffffeff`ffeeef
\eeacff^cfffdc\a^dadeffc\eda^`dddaaY\a_^b    1
```

- Sequence.txt: PF reads only. Illumina fastq sequence file:

```
@HWI-EAS216:1:2:1:2017#0/1
GCCATGCTCAGGAACAAAGAAACGCGGCACAGAATG
+HWI-EAS216:1:2:1:2017#0/1
a_aa`^aaaaa`_aa`YW`__a`__`___aa`____
```

- export.txt: 22 columns includes all reads and mapping info. Each read has one line.(quality encoding phred64)
- sorted.txt: 22 columns includes mapped reads only
- anomaly.txt (eland_paired only): contains one line for each read for which two halves of the read did not align with a nominal distance and orientation to each other (mine for structural variation information)

CASAVA Output Files

- CASAVA1.8

- FASTQ.txt

```
@NCI-GA3:98:66068AAXX:4:1:4494:949 1:N:0:
NGGCCGTCAGGAACGAGATGTAGATCTCTCCTGACA
+
#333285554@CCC@:<:<<@C@@C@@@@@C@@@@@C
```

  - The header line is interpreted as follows:
    @ <instrument-name>:<run ID>:<flowcell ID>:<lane-number>:<tile-number>:
    <x-pos>: <y-pos> <read number>:<is filtered>:<control number>:<barcode sequence>
  - Quality encoding is ASCII offset of 33, which is phred 33, same as sanger reads

- SAM/BAM
  - Header:

```
@HD     VN:1.0          SO:coordinate
@PG     ID:CASAVA       VN:CASAVA-1.8.2  CL:/opt/nasapps/stow/illumina/casava_v1.8.2/bin/configureBuild.pl --sortKeepAllReads --samtoolsRefFile /banas/
nextgen2/illumina/PROC/RefGenomes/hg19/ordered/hg19.fa --inSampleDir Project_xxxx_101560/Sample_112 --outDir Project_xxx_101560/Sample_112/BAM --
bamChangeChromLabel=UCSC --targets sort bam --wa
@SQ     SN:chr1         LN:249250621
@SQ     SN:chr2         LN:243199373
…
@SQ     SN:chr22        LN:51304566
```

  - Record:

```
NCI-GA3:98:66068AAXX:4:1:9207:948     16     chr15  86638786      42     36M    *     0     0     TGCCCTTAAGAGGAGAAAAGACAGACAAGGATATAN
CC@@C@@CC@@@2222@@@@@@C@CC55555)+0,#   AS:i:-1 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:35T0     YT:Z:UU

NCI-GA3:98:66068AAXX:4:1:9456:934     4     *     0     0     *     *     0
 NTACAAAAAAAATGTTTTTAAAAATTAACTGGGCATA   #330/8800022@@@@222@@@@@@@@@@@@@@@@@   YT:Z:UU
```

  - Quality encoding is phred33.

Interpret SAM Record

SAM record column field and description

- 1 QNAME String, Query template name
- 2 FLAG Int [0,216-1] bitwise FLAG
- 3 RNAME String, Reference sequence name
- 4 POS Int [0,229-1] 1-based leftmost mapping Position
- 5 MAPQ Int [0,28-1] Mapping Quality
- 6 CIGAR String, CIGAR string
- 7 RNEXT String, Ref. name of the mate/next segment
- 8 PNEXT Int [0,229-1], Position of the mate/next segment
- 9 TLEN Int [-229+1,229-1] observed template insert size
- 10 SEQ String , query template sequence
- 11 QUAL String, query template quality, ASCII of Phred-scaled base Quality+33

Interpret bitwise flag:

- http://picard.sourceforge.net/explain-flags.html

BAM File Manipulation and Visualization

- Samtools: http://samtools.sourceforge.net/samtools.shtml
- Picard: http://picard.sourceforge.net
- BAMtools: http://sourceforge.net/projects/bamtools/
- BEDtools: http://code.google.com/p/bedtools/
- IGV: http://www.broadinstitute.org/igv
- UCSC Genome Browser

Variant Calling Programs

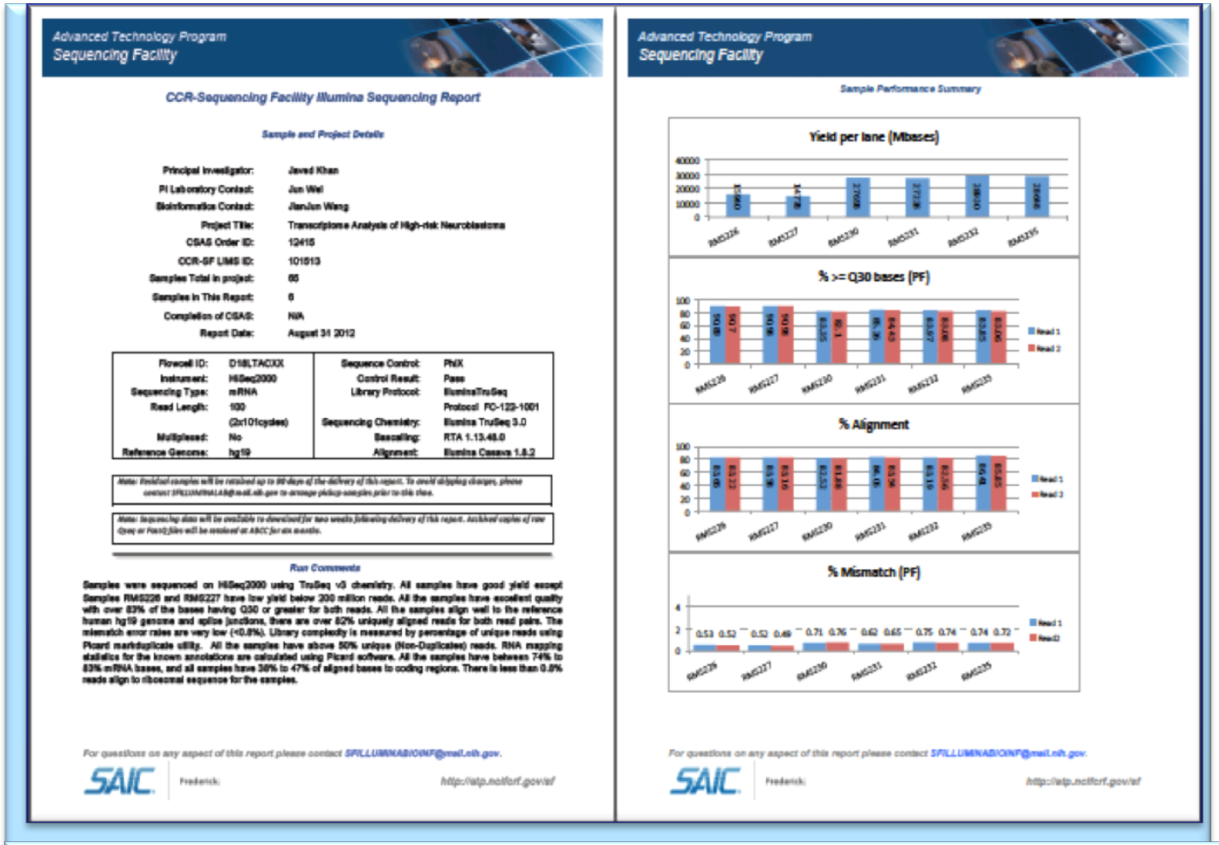
Tool	Input Format	Output Format	SNP	INDEL	Note
CASAVA	export	CSV	Y	Y	germline variant caller
SAMtools	BAM/ SAM	BCF, VCF	Y	Y	both germline and somatic variant caller
GATK	BAM	VCF	Y	Y	UnifiedGenotyper includes SNP and genotype caller
GATK SomaticIndelDetector		VCF	N	Y	somatic Indel caller
VarScan2	pileup, mpileup	VCF	Y	Y	both germline and somatic variant caller

Variant Call Files

CASAVA snps.txt file:

- 1 seq_name - Reference sequence label
- 2 pos - Sequence position of the site/snp
- 3 bcalls_use- Basecalls used to make the genotype call for this site
- 4 bcalls_filt- Basecalls mapped to the site but filtered out before genotype calling
- 5 Ref - Reference Base
- 6 Q(snp) - A Q-value expressing the probability of the homozygous reference genotype, subject to the expected rate of haplotype difference as expressed by the (Watterson) theta parameter *
- 7 max_gt - The most likely genotype (subject to theta, as above).
- 8 Q(max_gt) - A Q-value expressing the probability that the genotype is not the most likely genotype above (subject to theta).
- 9 max_gt|poly_site -The most likely genotype assuming this site is polymorphic with an expected allele frequency of 0.5 (theta is still used to calculate the probability of a third allele — i.e. the chance of observing two non-reference alleles).
- 10 Q(max_gt|poly_site) - A Q-value expressing the probability that the genotype is not the most likely genotype above assuming this site is polymorphic.
- 11 A_used - ‘A’ basecalls used
- 12 C_used - ‘C’ basecalls used
- 13 G_used - ‘G’ basecalls used
- 14 T_used - ‘T’ basecalls used

CCR-SF Data Delivery Report Files


Advanced Technology Program
Sequencing Facility

CCR-Sequencing Facility Illumina Sequencing Report

Sample and Project Details

Principal Investigator:	Javed Khan
PI Laboratory Contact:	Jun Wei
Bioinformatics Contact:	JianJun Wang
Project Title:	Transcriptome Analysis of High-risk Neuroblastoma
CSAS Order ID:	12415
CCR-SF LIMS ID:	101513
Samples Total in project:	66
Samples in This Report:	6
Completion of CSAS:	N/A
Report Date:	August 31 2012

Flowcell ID:	D18LTACXX	Sequence Control:	PhiX
Instrument:	HiSeq2000	Control Result:	Pass
Sequencing Type:	mRNA	Library Protocol:	IlluminaTruSeq Protocol FC-122-1001
Read Length:	100 (2x101cycles)	Sequencing Chemistry:	Illumina TruSeq 3.0
Multiplexed:	No	Basecalling:	RTA 1.13.48.0
Reference Genome:	hg19	Alignment:	Illumina Casava 1.8.2

Run Comments

Samples were sequenced on HiSeq2000 using TruSeq v3 chemistry. All samples have good yield except Samples RMS226 and RMS227 have low yield below 200 million reads. All the samples have excellent quality with over 83% of the bases having Q30 or greater for both reads. All the samples align well to the reference human hg19 genome and splice junctions, there are over 82% uniquely aligned reads for both read pairs. The mismatch error rates are very low (<0.8%). Library complexity is measured by percentage of unique reads using Picard markduplicate utility. All the samples have above 50% unique (Non-Duplicates) reads. RNA mapping statistics for the known annotations are calculated using Picard software. All the samples have between 74% to 83% mRNA bases, and all samples have 36% to 47% of aligned bases to coding regions. There is less than 0.8% reads align to ribosomal sequence for the samples.

For questions on any aspect of this report please contact SFILLUMINAB@mail.nih.gov.

SAIC Frederick http://atp.ncifcrf.gov/sf

Advanced Technology Program
Sequencing Facility

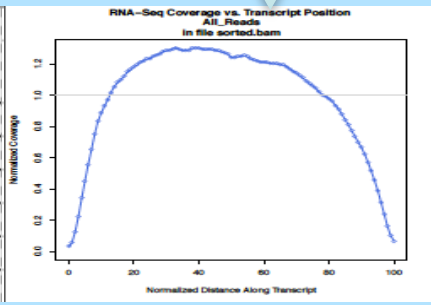
Sample Performance Summary

For questions on any aspect of this report please contact SFILLUMINAB@mail.nih.gov.

SAIC Frederick http://atp.ncifcrf.gov/sf

Custom sequencing and metrics reports produced in-house

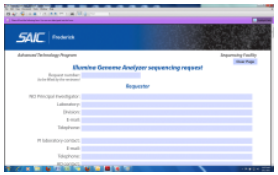
Sample	Sample Yield (Mbases)	Reads (PF)	Read1			Read2			% Unique reads(non-duplicates)	Library Size	PCT_CODING_BASES	PCT_UTR_BASES	PCT_INTRONIC_BASES	PCT_INTERGENIC_BASES	PCT_MRNA_BASES	PCT_USABLE_BASES
			% Align (PF)	% Mismatch Rate (PF)	% >=Q30 bases (PF)	% Align (PF)	% Mismatch Rate (PF)	% >=Q30 bases (PF)								
RMS226	15,940	159,401,324	83.66	0.53	90.89	83.22	0.52	90.7	66.5	321	38%	39%	11%	11%	78%	65%
RMS227	14,778	147,784,274	83.58	0.52	90.98	83.16	0.49	90.98	70.2	327	42%	32%	13%	13%	74%	61%
RMS230	27,698	276,973,628	82.52	0.71	83.35	81.88	0.76	82.1	49.3	323	45%	38%	7%	10%	83%	68%
RMS231	27,228	272,284,266	84.05	0.62	85.26	83.54	0.65	84.43	68.5	316	47%	31%	12%	10%	78%	65%
RMS232	28,930	289,298,514	83.19	0.75	83.97	82.56	0.74	83.08	53.2	316	42%	37%	9%	12%	79%	65%
RMS235	28,648	286,482,348	86.41	0.74	83.85	85.85	0.72	83.06	66.3	318	42%	33%	12%	13%	75%	65%



Steps to Bring in Projects to CCR–SF

Meet with SF group to discuss project scope and design

Sequencing proposal submission



CSAS submission

http:// ncifrederick.cancer.gov /rtp /csas /requestor/

Sample manifest

Please visit CCR-SF website

http://ncifrederick.cancer.gov/atp/genetics-and-genomics/sequencing-facility

Sample shipment

Post-SF IFX Options

- CCR-IFX Core – a bioinformatics core dedicated to supporting CCR labs bioinformatics requests.
    - Provide services for NGS data analysis
    - Pathway mapping and biological interpretation
    - miRNA and array CGH analysis
    - SNP and basecalling
    - Mutli-experiment data integration
  - http://ccrifx.cancer.gov/apps/site/analysis_support_process
- Other bioinformatics resources across the community
  - ABCC Bioinformatics Support Group:
    - http://isp.ncifcrf.gov/abcc/abcc-groups/bioinformatics
  - Bioinformatics@CCR
- BaseSpace (Illumina hosted cloud-based analysis platform)

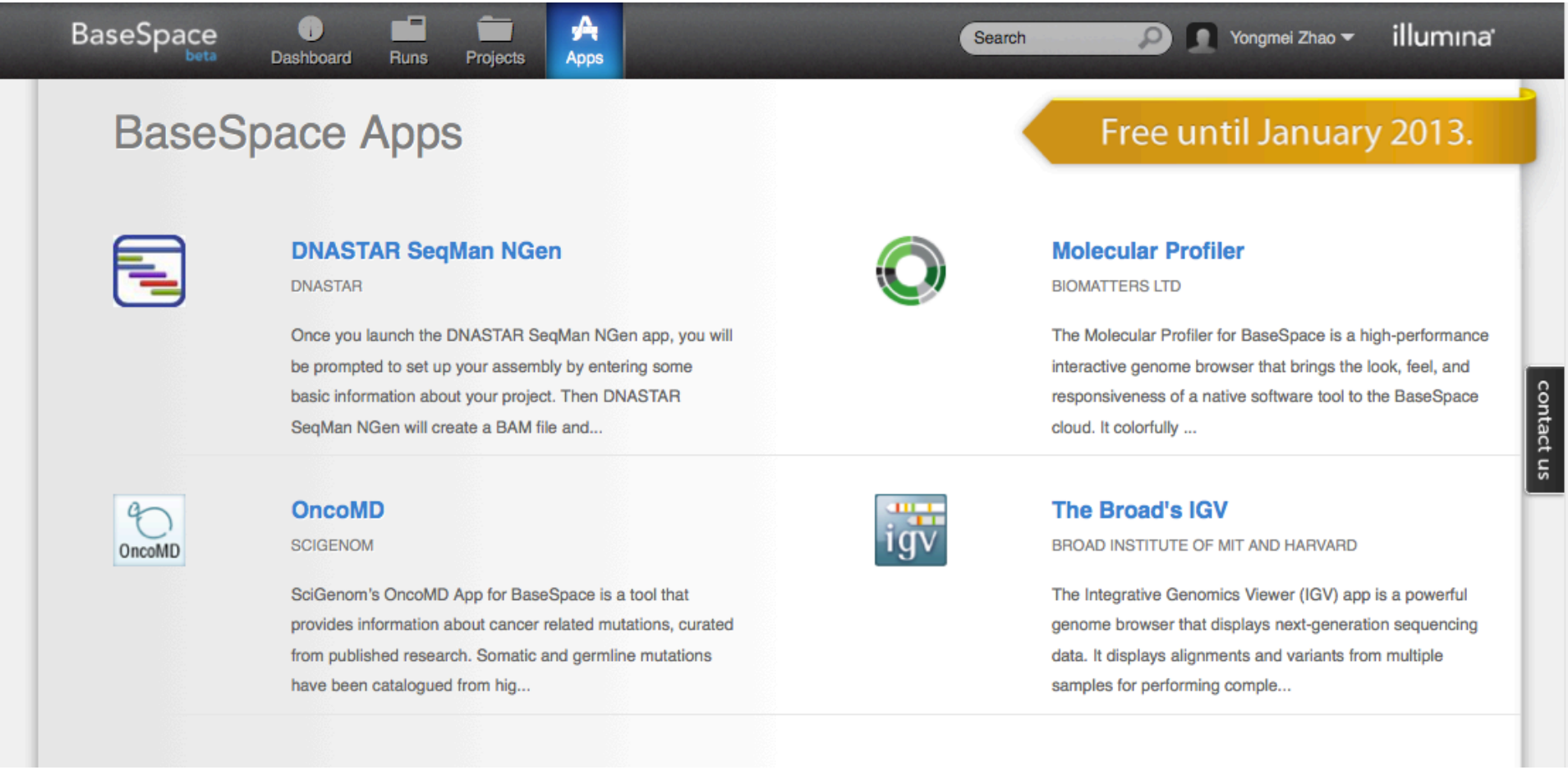
Illumina BaseSpace

https://basespace.illumina.com

BaseSpace Highlights

- Real-Time data upload and run monitoring(Currently support MiSeq, HiSeq1500 and HiSeq2500)
- Easy access to a growing collection of bioinformatics tools for QC, and downstream analysis and result visualization.
- Collaboration and data sharing on a global scale
- Enhanced security with data hosted on Amazon Web Services(AWS) that meet industry-accepted security standard.

BaseSpace Applications



https://basespace.illumina.com

Acknowledgments

CCR–SF

Bao Tran

Lab Team:
Jyoti Shetty
Yuliya Kriga
Yelena Levin
Tatyana Smirnova
Castle Raley

QC Team:
Jessica Dickens

IFX Team:
Yongmei Zhao
Shashi Ratnayake
Keyur Talsania

SAIC-Frederick, Inc.

Dwight Nissley
Jack Collins

ABCC BSG Group:
Robert Stephens
Ming Yi
Jigiu Shan

CCR-IFX Core:
Eric Stahlberg
Natalie Abrams

National Cancer Institute

Paul Meltzer
David Goldstein
Javed Khan
Louis Staudt
Gordon Hager
Robert Wiltrout

Bioinformatics Training and Education Program
Office of Science & Technology Partnerships (OSTP)