Ocimum sanctum – *Denovo* Whole Genome Sequencing

Genotypic Project ID:

SO_2054

For,
Dr. Vikrant Gupta,
Central Institute of Medicinal and Aromatic Plants,
Lucknow, India.

Genotypic Technology [P] Ltd.,
#2/13, Balaji Complex, 80 feet road, R.M.V. 2nd Stage, Bangalore-560094, INDIA
Phone: +91 80 40538202/ 8213; Fax: +91 80 40538222
Website: www.genotypic.co.in, **E-mail:** ngs.analysis@genotypic.co.in

Table of Contents

Bioinformatics Data Analysis

Module 1: Objective of the project

De-novo whole genome sequencing and assembly of *ocimum sanctum*.

1.1: Sample

Extracted genomic DNA from plant sample.

1.2: Sequencing platforms

Illumina HiSeq2000: 2 libraries and 454 GS FLX: 1 library.

1.3: Bioinformatics Data analysis tools

Edena v3.1, SSPACE v2.0, Misa v1.0, BLAT, Newbler

Module 2: Paired End Sequencing Using Illumina Chemistry

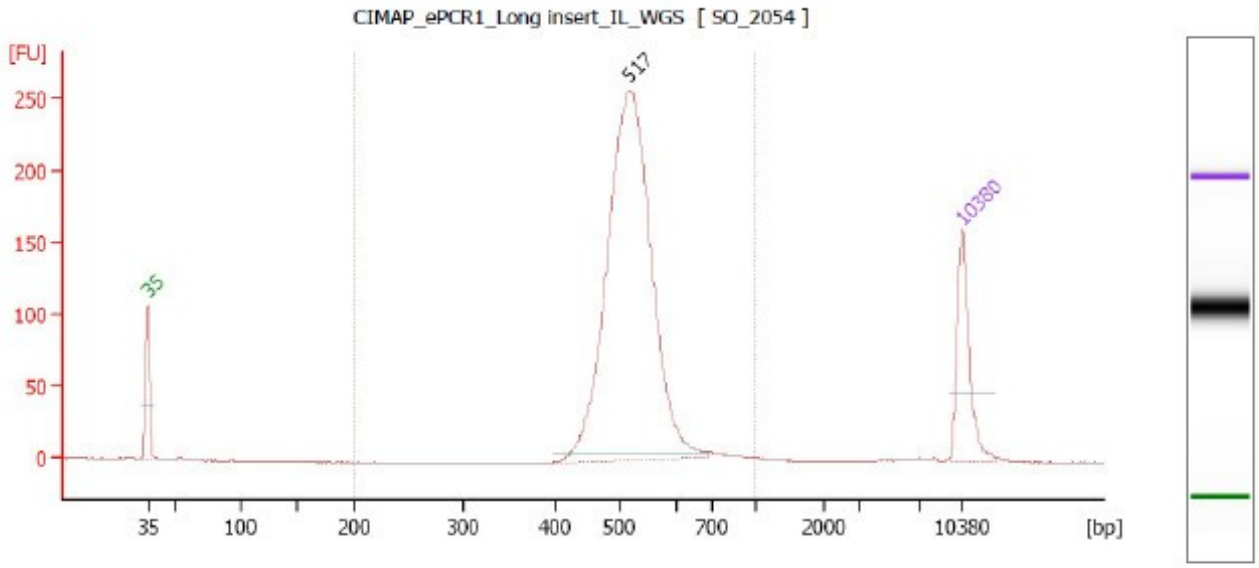
Standard fragment library represents a collection of regions of 300-500 bp insert length obtained from randomly fragmented genomic DNA. Individual single-stranded inserts function as template molecules during sequencing. Two ends of such inserts are sequenced for a length of 100 nucleotide each in a 100 PE sequencing run. Illumina chemistry detects bases using a sequencing by synthesis approach, where a DNA polymerase inserts fluorescent-tagged bases corresponding to the template molecules. This process is carried out in a massively parallel manner to enable the sequencing of millions of insert molecules simultaneously.

2.1: Methodology

~2 microgram of genomic DNA was made up to 50ul with nuclease free water (Ambion) and sonicated using covaris to obtain desired fragment length ranging between 150 to 600 bp. The resulting fragmented DNA was cleaned up using Agencourt Ampure XP SPRI beads (Beckman Coulter). The size distribution was checked by running an aliquot of the sample on Agilent High Sensitivity Bioanalyzer Chip.

Subsequently libraries for whole genome sequencing were constructed according to the TruSeq DNA library protocol outlined in “TruSeq DNA Sample preparation guide” (Part # 15005180; Rev. A ;Nov 2010). DNA was subjected to a series of enzymatic reactions that repair frayed ends, phosphorylate the fragments, and add a single nucleotide A overhang and ligate adaptors (Illumina’s TruSeq DNA sample preparation kit). Sample cleanup was done using Ampure SPRI beads. After ligation, ~500 – 600 bp for long insert, and 300-400bpfor short insert were size selectedon a 2% agarose gel and cleaned using MinElute column (QIAGEN). PCR amplification was done and cleaned up using Ampure SPRI beads. The prepared libraries were quantified using Nanodrop and validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent). (Refer Figure for Bioanalyzer profiles of amplified product ePCR1)

Figure 1. Bioanalyzer profile of amplified product (ePCR1) - Long Insert : (Avg. Insert Length is 524bp)



Overall Results for sample 9 : CIMAP_ePCR1_Long insert_IL_WGS

Number of peaks found:	1	Corr. Area 1:	1,458.1
Noise:	0.2		

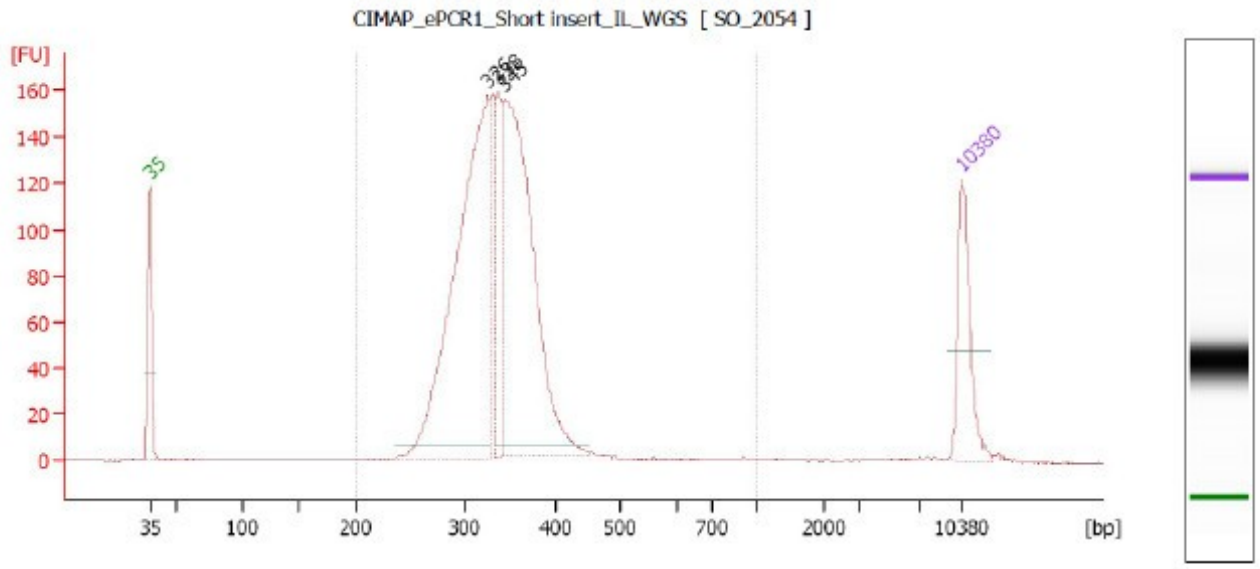
Peak table for sample 9 : CIMAP_ePCR1_Long insert_IL_WGS

Peak		Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	◀	35	125.00	5,411.3	Lower Marker
2		517	840.94	2,462.2	
3	▶	10,380	75.00	10.9	Upper Marker

Region table for sample 9 : CIMAP_ePCR1_Long insert_IL_WGS

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	1,458.1	98	524	9.9	848.35	2,475.0	

Figure 2. Bioanalyzer profile of amplified product (ePCR1) - Short Insert : (Avg. Insert Length is 340bp)



Overall Results for sample 10 : **CIMAP_ePCR1_Short insert_IL_WGS**

Number of peaks found:	3	Corr. Area 1:	1,774.9
Noise:	0.2		

Peak table for sample 10 : **CIMAP_ePCR1_Short insert_IL_WGS**

Peak	Size [bp]	Conc. [pg/µl]	Molarity [pmol/l]	Observations
1	35	125.00	5,411.3	Lower Marker
2	326	643.18	2,985.8	
3	338	108.97	488.3	
4	345	562.87	2,473.1	
5	10,380	75.00	10.9	Upper Marker

Region table for sample 10 : **CIMAP_ePCR1_Short insert_IL_WGS**

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
200	1,000	1,774.9	97	340	17.5	1,482.96	6,763.4	

2.2: Nanodrop values of prepared library

#	Sample ID	Nucleic Acid Conc.	Unit	260/280	260/230	Yield(ng)	NEXTflex Barcode
1	SO_2054_CIMAP_ePCR1_Long insert_IL_WGS	17	ng/µl	1.9	1.09	340	38
2	SO_2054_CIMAP_ePCR1_Short insert_IL_WGS	28.4	ng/µl	1.94	1.01	568	39

Comments: The library shows a peak at the range of 300-400bp for short insert library and 500-600bp for long insert library. The effective sequencing insert size is 180-280bp for short insert and 380-480bp for long insert; the inserts are flanked by adaptors whose combined size 120 bp. The libraries are suitable for 100PE sequencing on Illumina.

Once sequencing was completed, the raw data was extracted from the server using the proprietary Illumina pipeline software to obtain FASTQ files.

Module 3: Single-end Sequencing Using 454 GS FLX

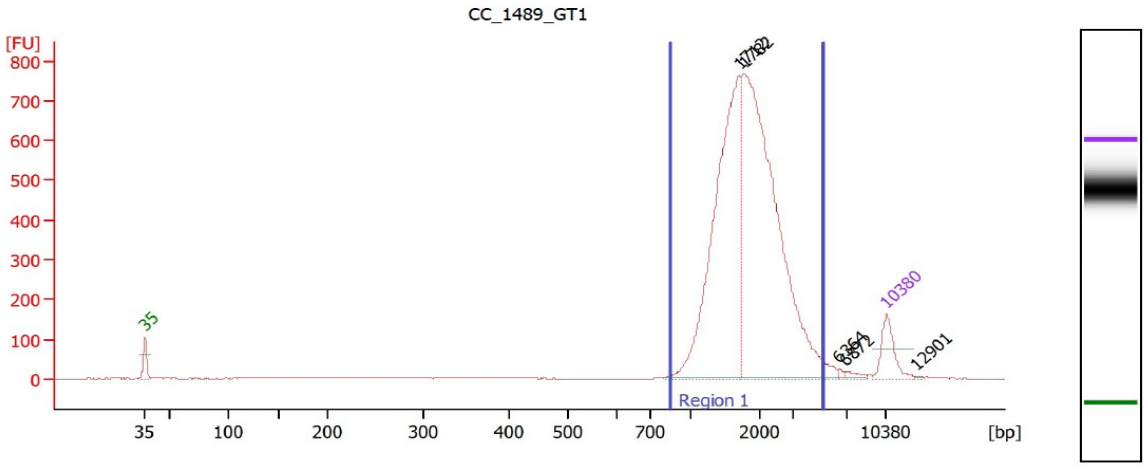
3.1: Materials

1. GS FLX Titanium Rapid Library Preparation Kit (Roche, #05 608 228 001)
2. Agencourt AMPURE XP beads (Beckman Coulter, #A63881)
3. MinElute PCR purification kit (QIAGEN, #)
4. High Sensitivity Bioanalyzer Chips (Agilent, #5067-4626)

3.2: Methods

Library for sequencing was constructed according to the Roche Rapid Library Preparation Method Manual (GS FLX+ Series – XL+, May 2011). Briefly, ~1 microgram of genomic DNA was made up to 100 ul with TE buffer and fragmented using a nebulizer. The resulting fragmented DNA was cleaned up using QIAGEN Minelute PCR purification kit. Subsequently, DNA was subjected to a series of enzymatic reactions that repair frayed ends, and ligates adaptors. After ligation, small fragments were removed using Agencourt Ampure SPRI beads. The prepared library was validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent). (Refer Figure 3 for Bioanalyzer profiles of amplified product ePCR1)

Figure 3: Bioanalyzer profile of amplified product (ePCR1)



Overall Results for sample 1 : CC_1489_GT1

Number of peaks found:	5	Corr. Area 1:	5,157.4
Noise:	0.5		

Peak table for sample 1 : CC_1489_GT1

Peak	Conc. [pg/µl]	Size [bp]	Molarity [pmol/l]	Observations
1	125.00	35	5,411.3	Lower Marker
2	864.34	1,712	765.1	
3	1,197.79	1,782	1,018.6	
4	5.33	6,364	1.3	
5	11.01	6,872	2.4	
6	75.00	10,380	10.9	Upper Marker
7	0.00	12,901	0.0	

Region table for sample 1 : CC_1489_GT1

From [bp]	To [bp]	Corr. Area	% of Total	Average Size [bp]	Size distribution in CV [%]	Conc. [pg/µl]	Molarity [pmol/l]	Color
853	5,216	5,157.4	95	1,964	33.4	2,036.34	1,695.8	

The library shows a peak at the range of 1.4 – 1.8 kb. The library is suitable for sequencing on 454 platform.

Module 4: Raw data QC

The Illumina HiSeq2000 paired end raw reads and 454 single-end reads were quality checked using Genotypic Pvt. Ltd., proprietary tool SeqQC[1].

Platform	Type of reads	Total number of raw reads
Illumina HiSeq2000	Paired end (101bp max)	224617107 pairs (45.37Gb of data)
454	Single End	643134(320.2Mb of data)

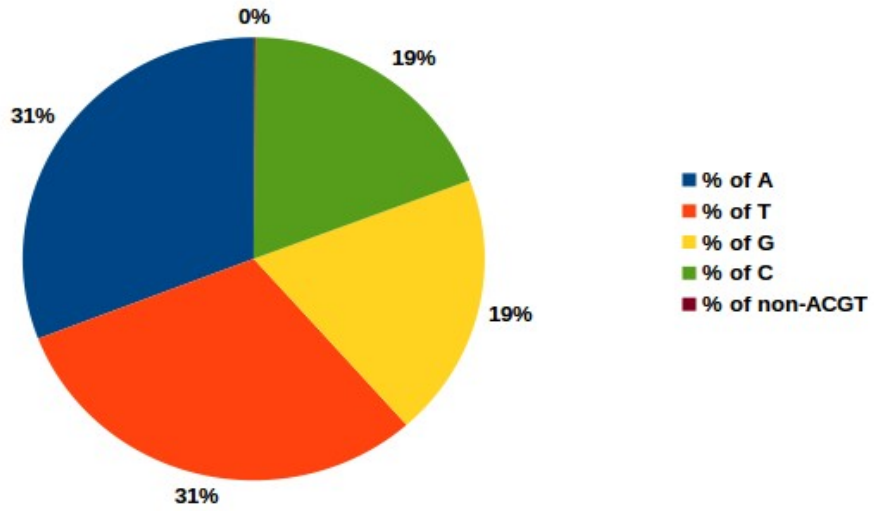
Total Data Sequenced = 45.69Gb (Illumina HiSeq2000 + 454)

4.1: GC content table of sequenced raw data

The GC content for Illumina as well as 454 reads is as follow

Sequencing technology	GC %
Illumina HiSeq2000	38.37
454	37.09

4.2: ATGC Composition of Illumina HiSeq2000 data



Description	Percentage
% of A	30.86
% of T	30.64
% of G	19.25

% of C	19.12
% of non-ACGT	0.10

Both of the Illumina libraries (short insert and long insert) have uniform read length of 101.

4.3 Read length Distribution of 454 GS FLX data

The 454 reads having varying read length ranging from, 40 to 999, i.e. 40 being the minimum read length and 999 as maximum read length. It has total number of 643134 reads.

Below is the read length distribution chart for 454 reads:

Range	Number of reads	Percentage of total reads
800-999	64810	10.08
600-799	210235	32.69
400-599	137998	21.46
200-399	107985	16.79
100-199	73890	11.49
99 or less	48216	7.5

4.4 Estimated coverage for each of the libraries

Library name	total bases	coverage(assuming 450Mb genome)
SO_2054_CIMAP_Long_insert_7	6284016990	13.96x
SO_2054_CIMAP_Long_insert	1932547736	4.29x
SO_2054_CIMAP_Short_insert_4	29717821658	66.03x
SO_2054_CIMAP_Short_insert	7438269230	16.52x

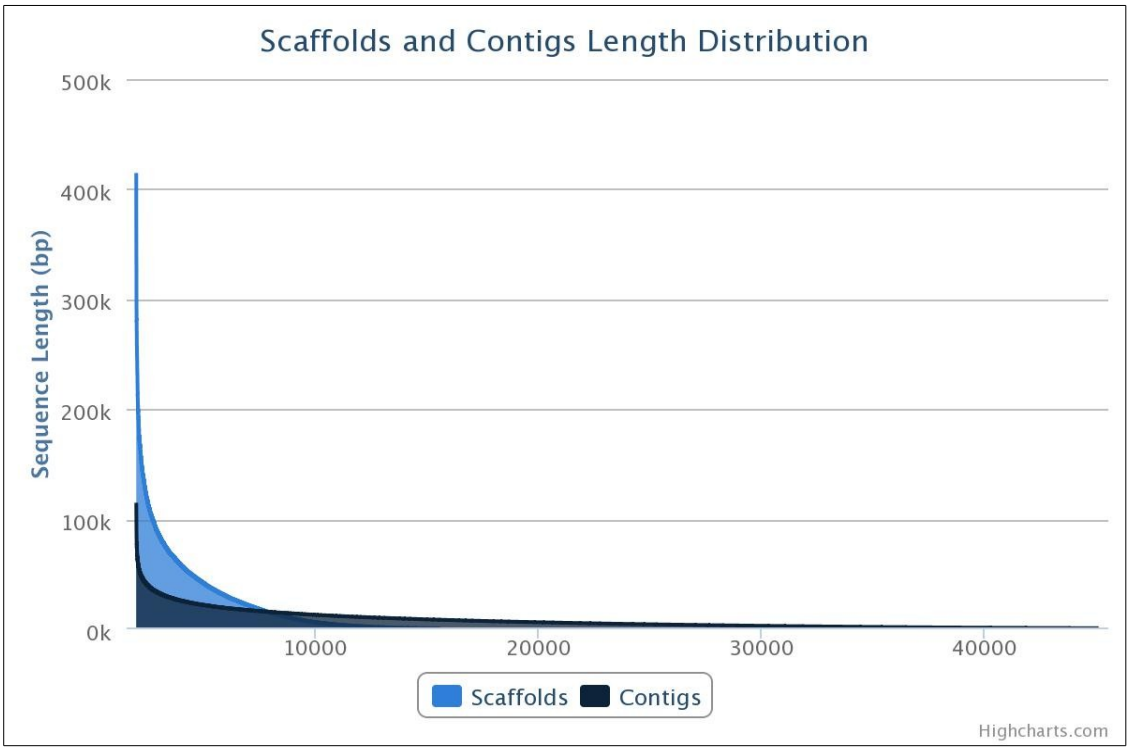
Module 5: De-novo assembly of sequenced genome data:

De-novo assembly of Illumina HiSeq2000 data was performed using Edena v3.1. Illumina paired-end data along with 454 GS FLX single end data were used for contig merging as a result of which scaffolds were generated. This scaffolding process was performed using the software SSPACE-2.03[3].

5.1 Assembly statistics of contigs and scaffolds based on length:

Description	Contigs	Scaffolds
Contigs Generated	107785	22776
Maximum Contig Length	115044	414711
Minimum Contig Length	147	200
Average Contig Length	3454	16984
Total Contigs Length	372395755	386828951
Total Number of Non-ATGC Characters	0	17898452
Percentage of Non-ATGC Characters	0	4.627
Contigs >= 1 Kb	43174	14791
Contigs >= 10 Kb	11594	7544
N50 value	12769	61854
N90 value	2071	12742

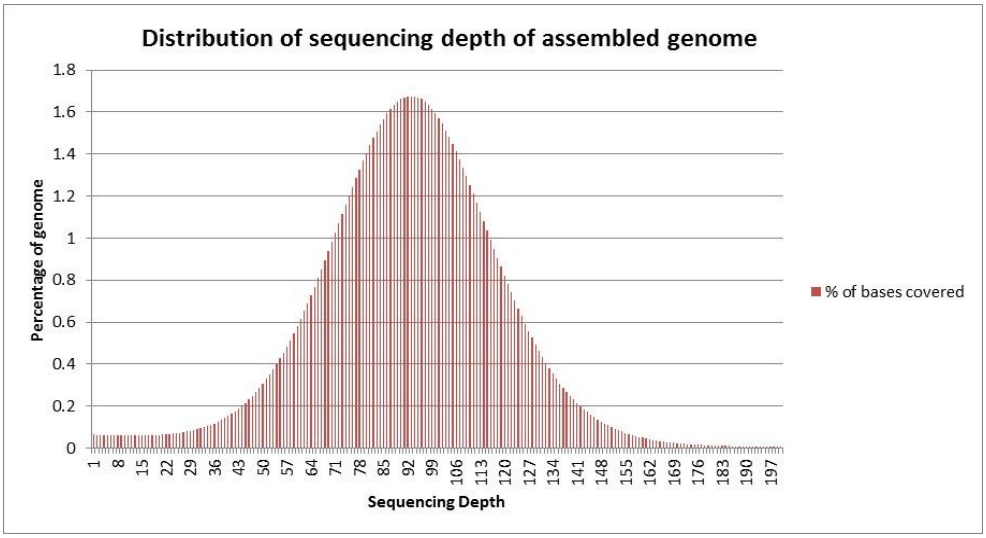
Comments on Assembly Statistics :
1. With the help of two Illumina library data (short insert and long insert) the assembly showed significant improvement in respect of n50.
2. The draft genome (scaffolds) resulted 22776 sequences which gives a total genome length of 386Mb.



The graph above shows the increase in sequence length and significant decrease in number of gaps as a result of scaffolfing.

5.2 Assessment of assembly quality.

Alignment of the raw data to the scaffolds have been caried out to calculate the sequencing depth of assembled genome. The graph below shows base wise read depth of scaffolds.



5.3 Evidence of genes on scaffolds

All of the publicly available mRNA, CDS and ESTs of Ocimum sanctum and Ocumum tenuiflorum were downloaded from NCBI. These were clustered using cd-hit-est to obtain 94 unigenes. These unigenes were aligned against scaffolds using BLAT. Query coverage cutoff 30% was applied. 76 out of 94 unigenes have found hits with the assembled scaffolds.

Module 6: Gene prediction and Annotation

Gene were predicted using Augustus from these scaffolds. Often it is the most accurate *ab initio* program. The predicted proteins were then mapped (BLASTP) with Uniprot all viridiplantae clade protein sequences. Blast hits were filtered with minimum 30% subject coverage and 30% percent identity criteria. Top hit from each of the query is considered further in this study.

Predicted gene annotation summary	Number
Total genes	85723
Total Annotated genes with UNIPROT	53480
Total Unannotated genes	32243

Module 7: Simple sequence repeat prediction (microsatellites)

Scaffolds were subjected to evaluate Simple Sequence Repeats

Description	Particulars
Total number of scaffolds examined	22776
Total size of examined sequences (Mb)	386.8Mb
Total number of identified SSRs	133236
Number of SSR containing scaffolds	11255
Number of SSRs with 1 unit	76064
Number of SSRs with 2 units	32293
Number of SSRs with 3 units	10939
Number of SSRs with 4 units	1285
Number of SSRs with 5 units	336
Number of SSRs with 6 units	434
Number of complex SSRs	11885

Module 8: References

1. SeqQC: Raja Mugasimangalam, Krishna Mohan, Vasanthan Jayakumar, Mohamed Ashick M, Kapila G and Vidya Niranjan: Rapid Quality Control of Next Generation Sequence Data 2013 (manuscript under preparation).
2. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. D.R. Zerbino and E. Birney. Genome Research 18:821-829.
3. SSPACE: Scaffolding pre-assembled contigs using SSPACE, Marten Boetzer1,2, Christiaan V. Henkel3, Hans J. Jansen3, Derek Butler1 and Walter Pirovano1.
4. Jared T. Simpson and Richard Durbin. Efficient construction of an assembly string graph using the FM-index. Bioinformatics Oxford journals. Vol:26, Iss:12, p367-373.
5. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
6. Altschul, S; Gish, W; Miller, W; Myers, E; Lipman, D (October 1990). "Basic local alignment search tool". Journal of Molecular Biology 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2. PMID 2231712.