An Integrated Approach to Transposon-Mediated Full Length cDNA Sequencing

Butterfield Y, MacDonald K, Stott J, Yang G, Smailus D, Griffith O, Guin R,
Barber S, Girn N, Lee D, Prabhu A, Tsai M, Schein J, Jones S, Marra M

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Canada's Michael Smith
Genome Sciences Centre
www.bcgsc.ca

Contribution to MGC

Number of sequences submitted:

Human:	5552
Mouse:	1256
Clawed frog:	522
Zebrafish:	in progress

Current Clone Totals

No clones submitted:
7330

Number of bases sequenced:
14 Mb

Number of reads:
251,637

Average cDNA sequence length:
1938 bp

1. Abstract

As a participant in the Mammalian Gene Collection (MGC) initiative (http://mgc.nci.nih.gov/)[1], we have generated accurate and complete sequence for **6,808** human and mouse genes. We have derived both computational and laboratory techniques to efficiently sequence these small insert clones. In addition to the publicly available software Phred, Phrap[2] and Consed[3], we have developed a number of programs to expedite the sequencing of the clones, and improve communication between the biochemistry and bioinformatics activities in our laboratory.

After sizing of clones and EST sequencing, results from BLAST and BLAT are used to confirm the identity of the clones and to check for problematic clones such as chimeras. The complete sequence for smaller sized cDNAs can sometimes be deduced from EST sequences alone.

For determining the full-length sequence of larger cDNAs, we make use of the Mu transposon. A number of different laboratory strategies are used to facilitate the sequencing of clones from various libraries and vector systems. In order to completely avoid repeated sequencing of vector, we have used Gateway technology (Invitrogen) when possible. Sequencing libraries are constructed containing up to 96 cDNA clones into which transposons are randomly inserted. Algorithms make use of insert sizing and DNA concentration information to divide clones into appropriate pools and to ensure that sequencing reads from these pooled libraries cover clones of non-uniform size evenly.

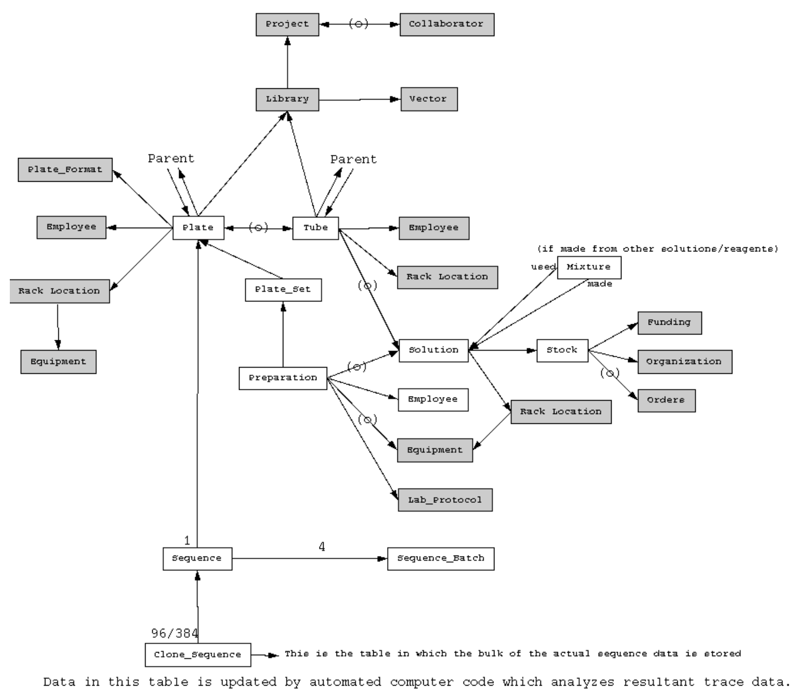
The EST sequences, sequences derived from transposons, and finishing reads for all the clones from a transposon pool are assembled together. Clones are automatically identified based on appropriate ESTs in the assembled contigs. We analyze sequence contigs computationally and automatically identify those that pass sequence integrity and quality checks and those that require further sequencing. We have also been able to analyze the transposon insertion profile and the effect on pooling sets of cDNA clones in various ratios.

2. Laboratory Information Management System (LIMS)

The GSC LIMS has been specifically designed for use in a genome sequencing laboratory. It has been developed to store detailed information related to standard sample preparation procedures as well as final sequence data.

By maintaining this detailed information in a highly structured format, there is the capability to perform a number of automatic procedures such as monitoring stock, error checking, and diagnostic analysis, generating real-time messages for users or regular email notifications to administrators. This helps to ensure the integrity of recorded data, and may prevent time-consuming errors by flagging them or identifying possible problems. Other project specific databases use the LIMS as a starting point for more in depth sequence analysis.

The database is implemented in MySQL and is comprised of about 75 tables. The largest table is the clone_sequence table having close to 1 million records and a size of 4 GB (roughly 4k/read x 1 million reads). The remainder of the database is 60 MB.



This is the table in which the bulk of the actual sequence data is stored

Data in this table is updated by automated computer code which analyzes resultant trace data.

Web front end

There is an interface which allows users to interact with the database via a barcode scanner (Fig. 2) during regular lab processes, and a sophisticated suite of report generating and data visualization tools on the web which provide lab administrators with the means to quickly and effectively evaluate results and monitor status on a regular basis (Fig. 1, 3).

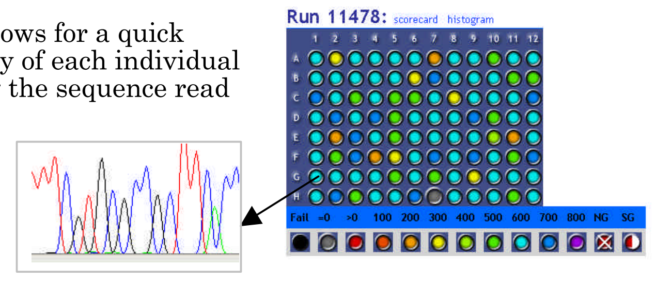

Run 11478: scorecard histogram
1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H
Fail =0 >0 100 200 300 400 500 600 700 800 NG SG

Figure 1. Plate view allows for a quick assessment of the quality of each individual sequence read. Clicking the sequence read brings up the trace.

Figure 2. Barcode scanner

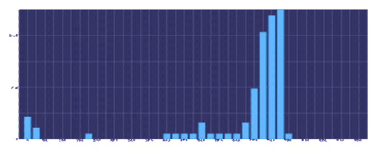
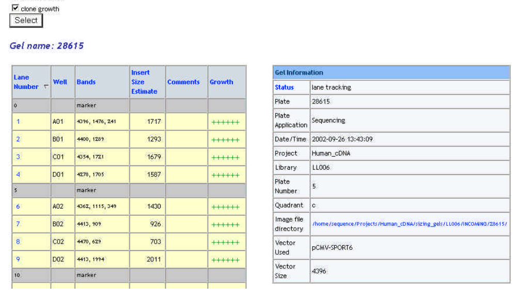


Figure 3. Phred Histogram showing number of reads vs the number of bases of Phred20 or higher. This histogram's peak is at 725bp.



Sizing Gels

Select Plate 28615 (LL006-5c) previous < > next
☑ marker lanes
☑ lane comments
☑ clone growth
Select

Gel name: 28615

Lane Number	Well	Bands	Insert Size Estimate	Comments	Growth
0		marker			
1	A01	4396, 1476, 241	1717		++++++
2	B01	4400, 1289	1293		++++++
3	C01	4354, 1721	1679		++++++
4	D01	4278, 1705	1587		++++++
5		marker			
6	A02	4362, 1115, 349	1430		++++++
7	B02	4413, 909	926		++++++
8	C02	4470, 629	703		++++++
9	D02	4413, 1994	2011		++++++
10		marker			

Gel Information	
Status	lane tracking
Plate	28615
Plate Application	Sequencing
Date/Time	2002-09-26 13:43:09
Project	Human_cDNA
Library	LL006
Plate Number	5
Quadrant	c
Image file directory	/home/sequence/Projects/Human_cDNA/sizing_gels/LL006/INCOMING/28615/
Vector Used	pCMV-SPORT6
Vector Size	4396

Figure 4. All cDNA clones are accurately sized. After lane-tracking, bands are automatically called with Bandleader[4] and results fed into the LIMS. Clone sizing and concentration information is required for the construction of transposon pools (see next section).

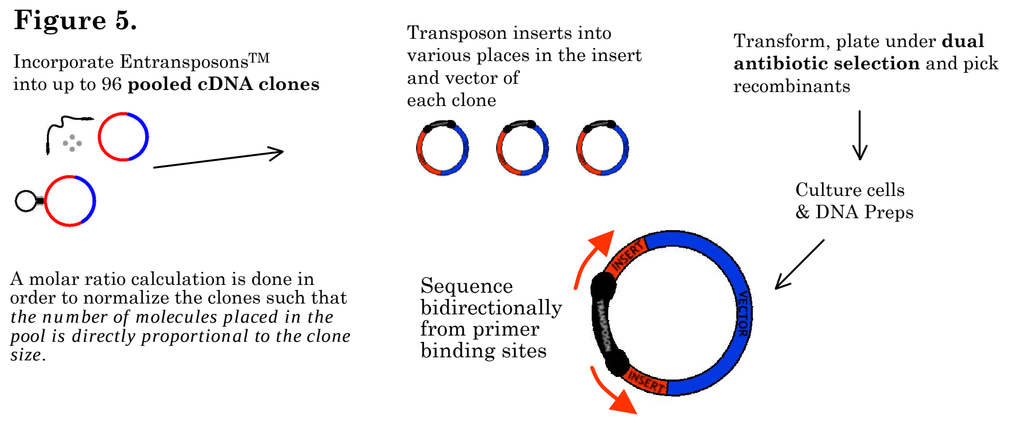
3. Full Length cDNA Sequencing

Overview of cDNA sequencing process[5]

The project relies heavily on the LIMS to track clone processing and sequencing through the pipeline. Clones are automatically sized with Bandleader and DNA concentration is also measured. This information is used to confirm final sequence assemblies and in molar ratio calculations.

After the generation of ESTs, clones are pooled together and "shotgun" reads derived from two transposon-specific primers are used to complete the cDNA sequence. DNA from remaining unfinished clones are either re-pooled into a second round of transposon mediated sequencing or are finished using directed reads from custom oligonucleotide primers if the contigs have smaller sequence gap. Remaining clones after full shotgun that do not meet specified finishing criteria are also subjected to primer directed sequencing.

Figure 5.



Incorporate Entransposons™ into up to 96 **pooled cDNA clones**

Transposon inserts into various places in the insert and vector of each clone

Transform, plate under **dual antibiotic selection** and pick recombinants

Culture cells & DNA Preps

Sequence bidirectionally from primer binding sites

A molar ratio calculation is done in order to normalize the clones such that *the number of molecules placed in the pool is directly proportional to the clone size.*

4. cDNA sequencing pipeline

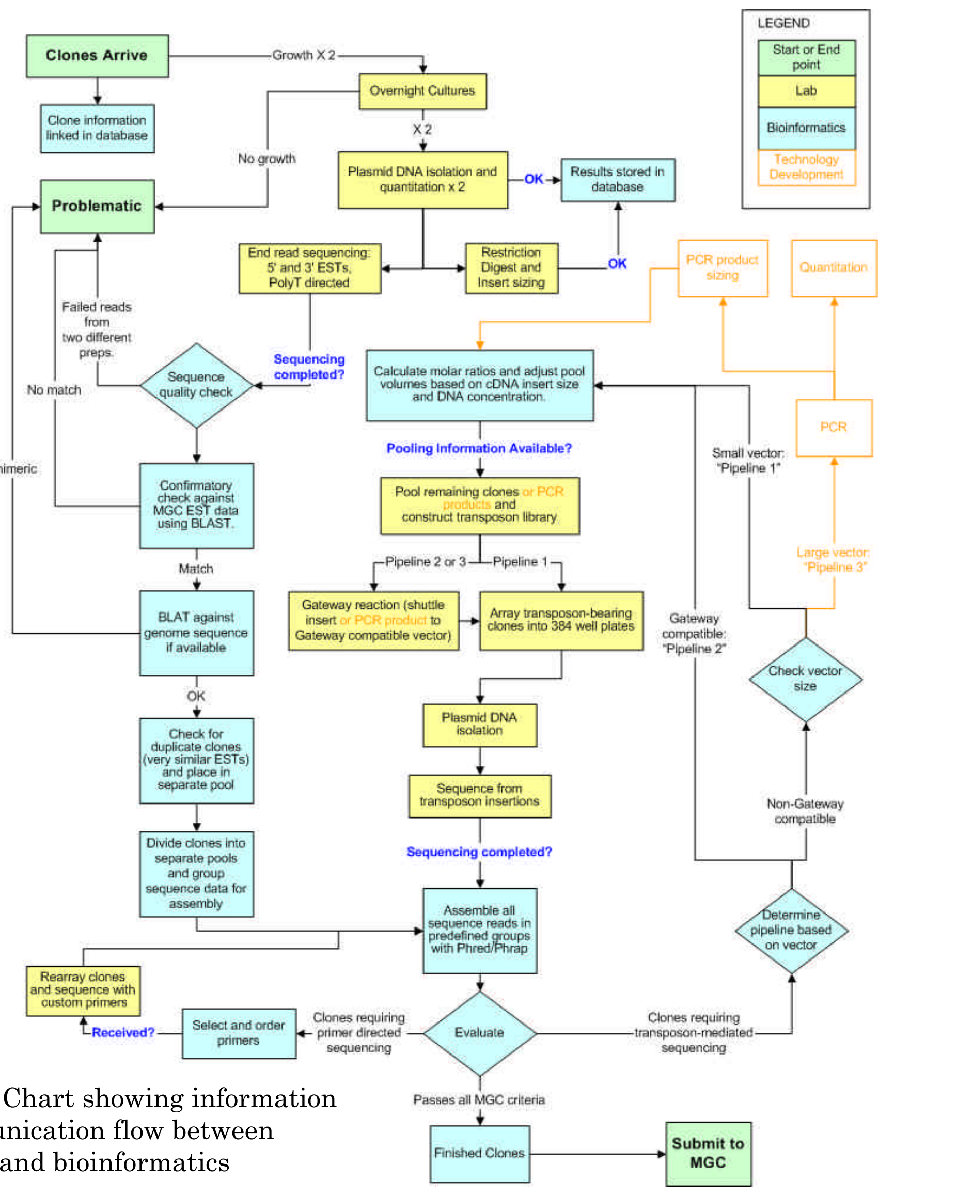

Clones Arrive
Growth X 2
Clone information linked in database
Overnight Cultures
X 2
No growth
Problematic
Plasmid DNA isolation and quantitation x 2
OK
Results stored in database
End read sequencing 5' and 3' ESTs, PolyT directed
Restriction Digest and insert sizing
OK
PCR product sizing
Quantitation
Failed reads from two different preps
No match
Sequence quality check
Sequencing completed?
Calculate molar ratios and adjust pool volumes based on cDNA insert size and DNA concentration
Chimeric
Pooling information Available?
Small vector: "Pipeline 1"
PCR
Confirmatory check against MGC EST data using BLAST
Pool remaining clones or PCR products and construct transposon library
Large vector "Pipeline 3"
Pipeline 2 or 3
Pipeline 1
Match
Gateway compatible "Pipeline 2"
BLAT against genome sequence if available
Gateway reaction (shuttle insert or PCR product to Gateway compatible vector)
Array transposon-bearing clones into 384 well plates
Check vector size
OK
Plasmid DNA isolation
Check for duplicate clones (very similar ESTs) and place in separate pool
Sequence from transposon insertions
Non-Gateway compatible
Divide clones into separate pools and group sequence data for assembly
Sequencing completed?
Assemble all sequence reads in predefined groups with Phred/Phrap
Determine pipeline based on vector
Rearray clones and sequence with custom primers
Received?
Select and order primers
Clones requiring primer directed sequencing
Evaluate
Clones requiring transposon-mediated sequencing
Passes all MGC criteria
Finished Clones
Submit to MGC
LEGEND
Start or End point
Lab
Bioinformatics
Technology Development

Figure 6. Chart showing information and communication flow between laboratory and bioinformatics

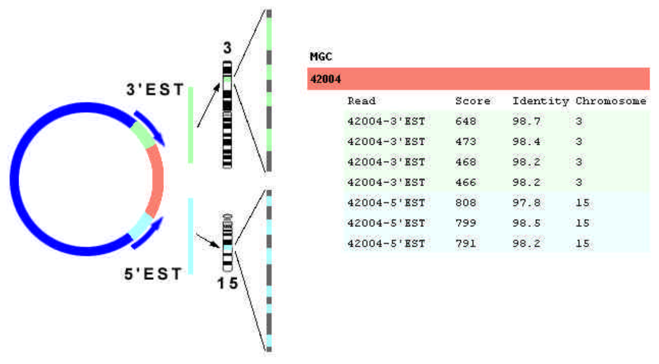
Figure 7. Summary of clones completed at different stages in the sequencing pipeline

	No. of clones	% Clones completed[a]	Total reads	Total sequence (Mb)[b]	Reads/kb	Average insert size (kb)	Average no. of reads/clone
Finished after one round of transposon sequencing	2149	58	77 837	4.17	18.7	1.9	36.2
Finished after second round of transposon sequencing	995	27	46 139	1.9	24.3	1.9	46.4
Finished with directed primer reads	501	14	14 240	0.97	14.7	1.9	28.4
Finished with end reads only[c]	50	1	375	0.04	10.6	0.8	7.5
Total	3695	100	138 591	7.06	19.6	1.9	37.5

[a] Percentage of total number of clones analyzed.
[b] Finished sequence.
[c] EST and oligo(dT) reads. Fifty-eight percent of the clones were completed with one round of transposon-mediated sequencing in pools of up to 96 clones. The remaining clones were placed into a secondary pool and subjected to a second round of transposon-mediated sequencing. A further 27% of the clones were completed after this step. A small fraction (1%) of clones were completed with only end sequence reads. Other clones (14%), required primer walking.

5. Informatics Tools

A number of perl scripts and modules process the data through the pipeline (Fig. 6). For example, before transposon sequencing, the paired ESTs for each clone are automatically analyzed to detect potential chimeras (Fig. 8).



MGC
42004

Read	Score	Identity	Chromosome
42004-3'EST	648	98.7	3
42004-3'EST	473	98.4	3
42004-3'EST	468	98.2	3
42004-3'EST	466	98.2	3
42004-5'EST	808	97.8	15
42004-5'EST	799	98.5	15
42004-5'EST	791	98.2	15

Figure 8. A clone is classified as chimeric if there are no matches to the same chromosome from both ESTs. Clone 42084 has been detected as chimeric because each clones respective 5' and 3' ESTs match to different human chromosomes using BLAT[6]. No further sequencing is done on these clones.

Informatics Tools

Transposon, EST and custom oligo sequence data are assembled using Phrap and results are stored in a separate MySQL cDNA database (Fig. 9). Scripts examine the assembly in greater detail than what Phrap itself provides such as read coverage along the length of the assembly and the incorporation of EST reads used to identify the clone.

Figure 9. Portion of cDNA database schema. The database contains information from phrap assemblies relevant to the automated completion of cDNA sequences. Application queries also link this database to the GSC LIMS to obtain sequence and clone specific information.

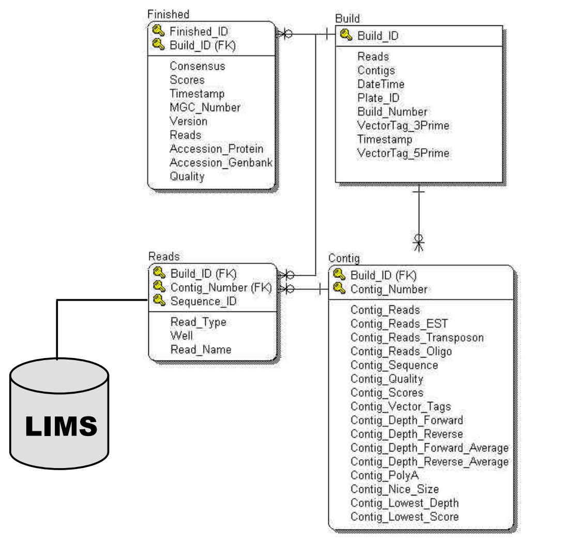
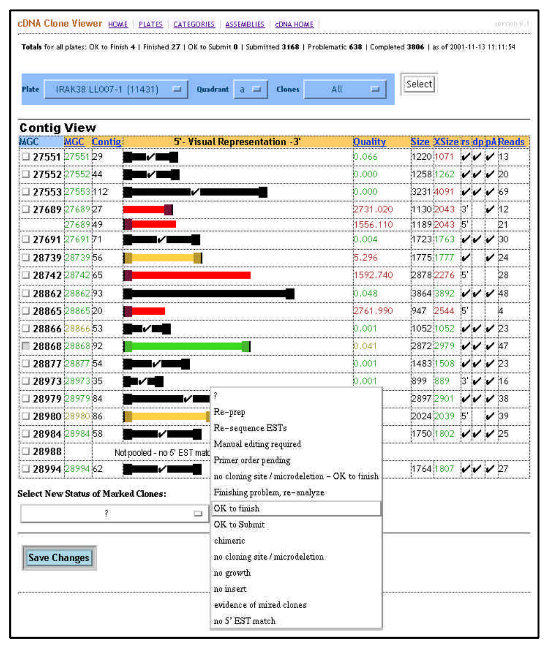


Figure 10. Web application for viewing progress of cDNA sequencing



cDNA Clone Viewer HOME | PLATES | CATEGORIES | ASSEMBLIES | cDNA HOME |

Totals for all plates: OK to Finish 4 | Finished 27 | OK to Submit 0 | Submitted 3168 | Problematic 638 | Completed 3806 | as of 2001-11-13 11:11:54

Plate IRAK38 LL007-1 (11431) Quadrant a Clones All Select

Contig View

MGC	MGC	Contig	5'- Visual Representation -3'	Quality	Size	XSize	rs	dp	pA	Reads
□ 27551	27551	29		0.066	1220	1071	✓	✓	✓	13
□ 27552	27552	44		0.000	1258	1262	✓	✓	✓	20
□ 27553	27553	112		0.000	3231	4091	✓	✓	✓	69
□ 27689	27689	27		2731.020	1130	2043	3'		✓	12
	27689	49		1556.110	1189	2043	5'			21
□ 27691	27691	71		0.004	1723	1763	✓	✓	✓	30
□ 28739	28739	56		5.296	1775	1777	✓		✓	24
□ 28742	28742	65		1592.740	2878	2276	5'			28
□ 28862	28862	93		0.048	3864	3892	✓	✓	✓	48
□ 28865	28865	20		2761.990	947	2544	5'			4
□ 28866	28866	53		0.001	1052	1052	✓	✓	✓	23
□ 28868	28868	92		0.041	2872	2979	✓	✓	✓	47
□ 28877	28877	54		0.001	1483	1508	✓	✓	✓	23
□ 28973	28973	35		0.001	899	889	3'	✓	✓	16
□ 28979	28979	84			2897	2901	✓	✓	✓	38
□ 28980	28980	86			2024	2039	5'		✓	39
□ 28984	28984	58			1750	1802	✓	✓	✓	25
□ 28988			Not pooled - no 5' EST matc							
□ 28994	28994	62			1764	1807	✓	✓	✓	27

Select New Status of Marked Clones: ?

Save Changes

?
Re-prep
Re-sequence ESTs
Manual editing required
Primer order pending
no cloning site / microdeletion – OK to finish
Finishing problem, re-analyze
OK to finish
OK to Submit
chimeric
no cloning site / microdeletion
no growth
no insert
evidence of mixed clones
no 5' EST match

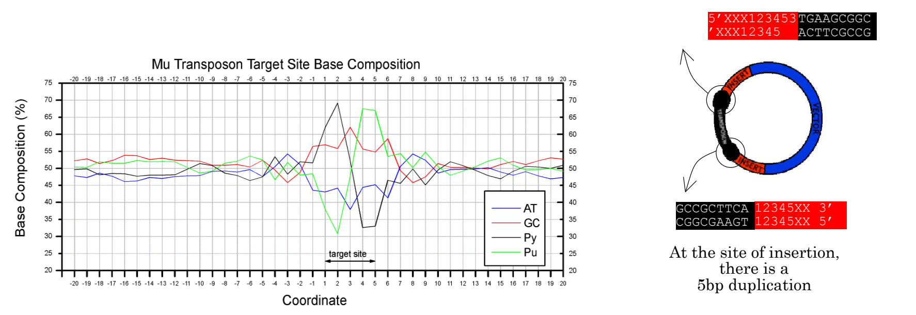
This database allows for the quick assessment of contigs and identification of clones; web tools facilitate the automatic visualization of the required criteria checks and *finishing of clones without having to look at each individual contig* (Fig. 10). Information such as clone and assembly sizes, quality, the number of reads, and various sequence integrity checks are available.

Consed is used where necessary when editing and finishing is required but the majority of the clones full length sequence have been determined with automated approaches alone.

The database and web tools are written in Perl modules and easily portable to other cDNA sequencing projects. The same code base is also used for EST clustering projects.

6. Transposon Based Sequencing Analysis

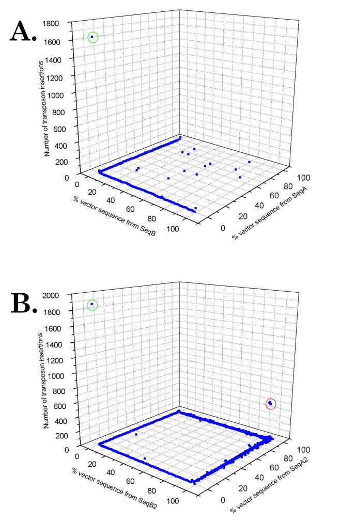
We have also developed software to analyse in greater detail the insertion profile of Mu transposon (Fig. 11) and to compare variations of sequencing with this method (Fig. 12).



At the site of insertion, there is a 5bp duplication

Figure 11. Insertion profile of the Mu transposon in the insert. An analysis of the insert region that spans this target site describes a consensus sequence preference for Mu transposon insertion. The insertion site displays a symmetry that includes the target site consisting of pyrimidines followed by purines as shown.

Figure 12. Two transposon sequencing approaches. Individual transposon insertions plotted as a function of the percentage of vector bases generated from each of the transposon-primed sequencing reads. The x- and y-axes indicate the percentage of vector sequence derived from each of the reads, and the z-axis indicates the number of insertions with a particular proportion of vector derived sequence from each read.



A. GeneJumper™-Gateway™ Sequencing. Using this system, the insert containing the transposon is shuttled into a new vector. This eliminates sequencing from clones where the transposon inserted into the vector.

B. Standard transposon sequencing methodology.

Acknowledgements

funding | National Cancer Institute

thanks | Staff of the BC Cancer Agency, Genome Sciences Centre

references |

1. Strausberg, B., et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *PNAS*. Dec 24;99(26):16899-903
2. Ewing, B. and Green, P., 1998. *Genome Research*. 8:186-194.
3. Gordon, D., et al. 1998. Consed: A graphical tool for sequence finishing. *Genome Research*. 8:195-202
4. Fuhrmann, Dan et al. 2001. *Automated Image Analysis for DNA Fingerprinting* (unpublished)
5. Butterfield, Y. et al. 2002 *Nucleic Acids Research*. Vol. 30, No. 11:2460-2468
6. Kent, W.J., 2002 *Genome Research*. Vol. 12:656-664