Transposon-Mediated cDNA Sequencing

at the British Columbia Cancer Agency, Genome Sciences Centre

D. Smailus, J. Asano, Y. Butterfield, N. Girn, R. Guin, M. Krzywinski, S. Lee, K. MacDonald, T. Olson, P. Pandoh, P. Saeedi, U. Skalska, L. Spence, J. Stott, S. Taylor, K. Teague, G. Yang, J. Schein, S. Jones and M. Marra.

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

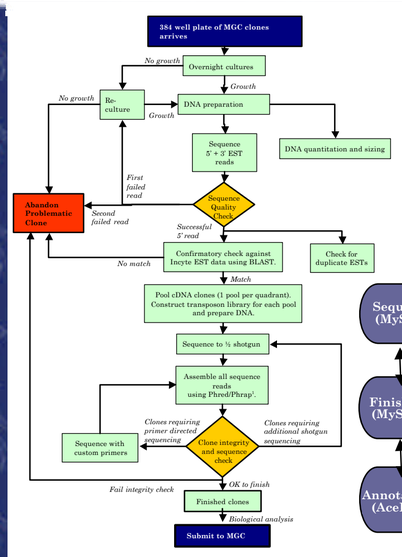
Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

Abstract

We have developed an efficient, high-throughput method for accurate DNA sequencing of entire cDNA clones through our participation in the NCI-sponsored Mammalian Gene Collection. Sequencing is accomplished though the insertion of Mu transposon into cDNAs, followed by sequencing reactions primed with Mu-specific sequencing primers. Transposon insertion reactions are not performed with individual cDNAs but rather on pools of up to 96 clones. Accurate clone insert size and DNA quantitation data are used to ensure proportional representation of each cDNA clone in the pool. This pooling strategy reduces the number of transposon insertion sequencing libraries that would otherwise be required, reducing the costs and enhancing the efficiency of the transposon library construction procedure. Sequences are assembled using Phred, Phrap, and Consed to yield the full-length cDNA sequence, with sequence editing and other sequence finishing activities performed as required to resolve sequence ambiguities. We are currently in our second year of the MGC project and have used the method to generate more than 7.5 Mb of finished sequence from 3,956 candidate full-length cDNAs. Analysis of 22,785 sequenced Mu transposon insertion events revealed a weak sequence preference for Mu insertion. However, the insertion pattern deviates only slightly from random and does not adversely affect the efficacy of our method. A detailed description of our transposon-mediated sequencing methodology and analysis of Mu transposon insertion events will be presented.

Pipeline Flow Chart



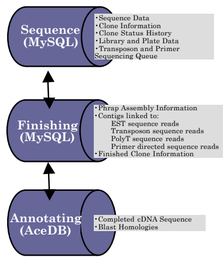
- The flow chart shows the processing of clones from the Mammalian Gene Collection.
- After the generation of ESTs, Shotgun reads are derived from two transposon-specific primers. DNA from remaining unfinished clones are either re-pooled into a second round of transposon mediated sequencing or are finished using directed reads from custom oligonucleotide primers if the contigs have smaller sequence gaps. Remaining clones after full shotgun that do not meet MGC specified finishing criteria are also subjected to primer directed sequencing.
- Problematic clones are as defined by MGC guidelines, and are removed from the pipeline.
- The pipeline involves data flow between the sequencing laboratory and bioinformatics and relies on a number of databases and programs.


Database web front-end


Graphical database schema view


Example of table relationships

Databases



Using the database, the entire process of DNA preparation, sequencing, data analysis and storage are tracked and stored. All aspects of run conditions and analysis are kept in database tables to facilitate locating trends in performance, keeping quality control and compiling statistics.

DNA Isolation for EST Sequencing

DNA is isolated using an alkaline lysis procedure modified from that of Marra et al, 1999.
384-well MGC glycerol stock from Christa Prange (LLNL) **http://mgc.ncl.nih.gov/**

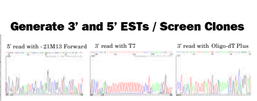
All plasticware, reagents, equipment and processes are tracked by barcodes. 384-well glycerol clone source plates shipped from Lawrence Livermore National Laboratory (LLNL) are received and logged into the database. Cultured cells are pelleted and wells that do not show evidence of bacterial cell growth ("no-grows") are documented in database. Quadrants containing no-grows are re-inoculated to confirm clones are truly non-viable and those failing to grow a second time are abandoned.

DNA Quantitation, EST Analysis and Clone Sizing

DNA purifications are quantitated in a 96-well spectrophotometer and the data stored for automatic retrieval.

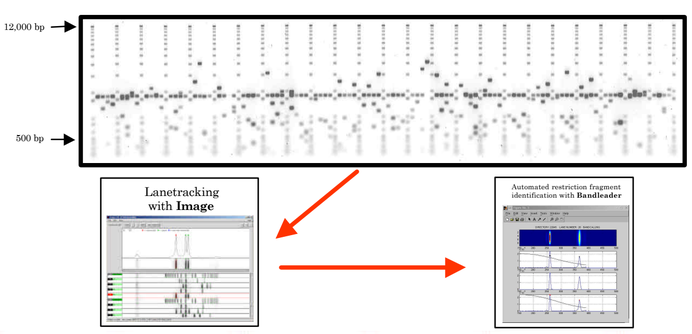
plate6160 LL005-24D 010323		218.98007 Avg ng/ul		37.76902 Std Deviation		8.8163 Minimum		307.04 Maximum			
	1	2	3	4	5	6	7	8	9	10	11	12
A	219.96	202.5	277.81	198.69	272.88	227.16	237.96	208.12	238.87	239.54	253.76	254.77
B	231.6	212.04	187.06	201.53	221.04	268.92	240.61	229.06	207.53	196.64	256.59	226.59
C	242.12	257.88	273.46	221.43	278.79	196.59	195.58	181.38	200.82	307.04	219.31	212.04
D	204.61	194.04	223.56	203.65	190.59	214.96	254.12	224.82	171.88	199.06	226.87	237.6
E	209.72	172.65	8.8163	187.41	197.65	174.24	195.18	166.32	179.67	192.96	207.17	235.06
F	228.91	221.29	202.78	268.56	245.52	248.12	185.29	215.29	296.28	228.96	181.73	206.08
G	255	206.64	192.49	277.71	227.76	253.47	192.12	276.48	189.92	210.12	172.87	154
H	186.35	212.62	263.52	231.12	229.59	234	202.32	210.24	241.35	215.65	220.6	271.06

Generate 3' and 5' ESTs / Screen Clones



5' and 3' reads for each clone are generated by Big Dye Terminat or cycle sequence analysis using –21M13 Forward primer, T7 and Oligo-dT Plus primers respectively on ABI 3700 DNA analyzers. We currently perform 0.25X chemistry in 4ul reaction volumes containing approximately 45ng of DNA. Successful reads are checked against Incyte or Agencort data and MGC IDs. Clones with failed reads are re-cultured; DNA isolated, sequenced and checked. Clones failing MGC ID checks twice are abandoned.

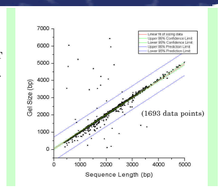
384 *Eco*RI restriction digested clones are sized by loading onto four 121-lane combs embedded into a single agarose gel (96 samples and 25 marker lanes per comb). The gel contains a DNA marker in every fifth lane. After electrophoresis, the gel is stained with SYBR Green and an image of the gel is collected on a Moleculare Dynamics Fluorimager.




Dr. Michael Smith
1932-2000
Founding Director

EST Analysis

- Analysis of data that are generated at various stages in the pipeline allows for both manual and automatic assessments of each clone.
- The first step is to confirm the identity and quality of clones that we have received by matching the EST reads against existing 5' EST sequences generated from the same clones by Incyte Inc. Clones that do not result in a positive match are automatically removed from the pipeline and reported as problematic.
- Failure to generate a confirmatory 5' EST read can result from no grows or cross-contaminated wells. The analyses are summarized in varying levels of detail on our cDNA web pages.
- Also, ESTs are checked for matches to other ESTs from the same quadrant in order to eliminate duplicates.
- Clone sizes are determined by restriction enzyme analysis and band called using Image (www.sanger.ac.uk/Software/Image) and Bandleader (Fuhrmann, Dan et al. 2001. *Automated Image Analysis for DNA Fingerprinting* (unpublished).
- The accuracy of the sizing is exemplified by the graph (shown at right) that compares the completed sequence length against size as determined by the gel.



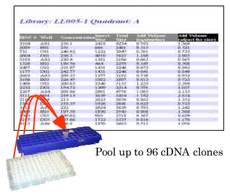
Various web pages show the status and location of clones, EST sequence qualities and Blast results.

IRAL42 (LL005-9A(15984) 96 clones selected

MGC	Location	Status	Conc.	Gel (bands)	Self Match Score	Best Incyte Match	Best 5' Score
39881	b-H05 (016)	Submitted	G	401.65 1181 (1793,1203)	0	29081	0
39894	b-A02 (A04)	Submitted	G	057.35 612 (1794,628)	0	29094	0
39895	b-H05 (016)	Submitted	G	269.59 2316 (2038,1793)	0	29095	0
39880	b-B05 (C10)	No cloning site / microdeletion		251.65 2270 (3565)	0	29100	0
39885	b-A11 (A22)	Submitted	G	209.39 1090 (1793,923,289)	0	29105	0
39882	b-G06 (A07)	Chimeric		195.6 1501 (1794,1281,307)	5e-126	29106	5e-126
39887	b-C08 (E16)	No 5' EST match		230 2463 (1939,1793,573)	4.7	29151	2e-06
39914	b-G05 (A01)	No 5' EST match		262.72 26 (1641)		29151	2e-06
39916	b-E03 (06)	Submitted	G	252.82 1509 (1794,1538)	0	29116	0

Pool cDNA Clones and Add Entransposons™

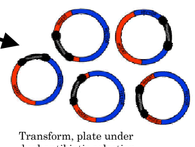
Molar Ratio Calculator:



Pool up to 96 cDNA clones

Two oligonucleotides that prime from either end of the randomly incorporated Entranceposons are used to shotgun sequence pooled clones. Total number of transformed clones to sequence for each shotgun library is calculated automatically at 11 reads/kb of target DNA ("half shotgun").

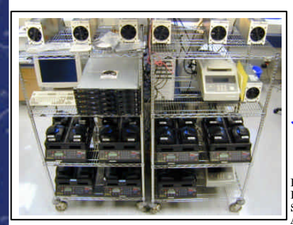
Incorporate Entransposons™ into pooled cDNA clones

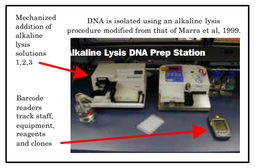


Transform, plate under dual antibiotic selection and pick recombinants

Transposon-Mediated cDNA Sequencing Pipeline

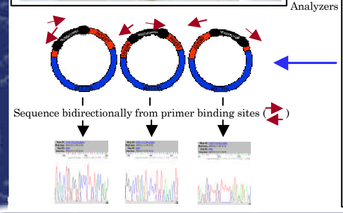
Transform DNA, plate under dual antibiotic selection and pick recombinants using QPix2


96 & 384-well Cycle Sequencing



Load DNA Sequence Analyzers

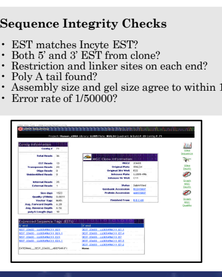

Six custom nodes share 1 user interface
Pharmacia Biosciences MegaBACE1000 & 4000 & ABI 3700 Sequencers

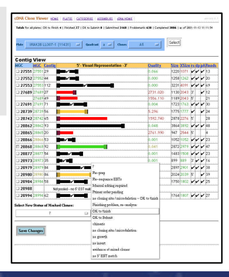


Web Tools

Sequence Integrity Checks

- EST matches Incyte EST?
- Both 5' and 3' EST from clone?
- Restriction and linker sites on each end?
- Poly A tail found?
- Assembly size and gel size agree to within 10%?
- Error rate of 1/50000?





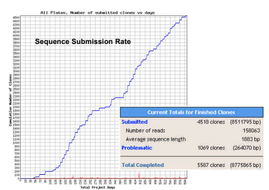
- Contig assembly information from the Phrap output file for each build is parsed into an SQL Finishing database.
- This database allows for the quick assessment of contigs and identification of clones.
- Web tools shown at right and below facilitate the automatic visualization of the required checks and *finishing of clones without having to look at each individual contig*.
- Information such as clone and assembly sizes, quality, the number of reads, and various sequence integrity checks are available.
- The interface also allows for manually changing the status of clones when needed. Other status changes are automatically made by various scripts.

Statistics

Stage	Total Clones	Fraction
Received (14 plates)	5376	100.0
Detected as problematic	524	9.7
Finished without transposon sequencing	176	3.3
Subjected to transposon sequencing	4676	87.0

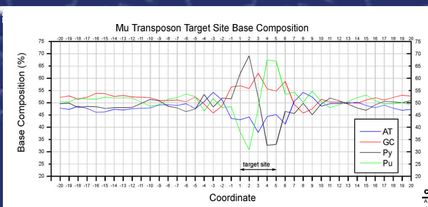
Transposon Pooling Success	Total Clones	Fraction	Total Reads	Mb	Reads/Kb	Avg insert Size (kb)	Avg # reads / clone
Finished after one round of transposon sequencing	2635	56.4	93548	5.1	18.3	1.9	36
Finished after second round of transposon sequencing	1251	26.8	49531	2.1	23.6	1.7	40
Finished with directed primer reads after transposon sequencing	712	15.2	18424	1.3	14.2	1.8	26
1 round	439						
2 rounds	158						
3 or more rounds	115						
Remaining	78	1.7					
Total	4676	100.0	161503	8.5	19.0	1.8	36

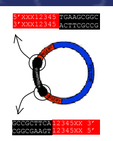
- A summary of the number of clones completed in the two shotgun mediated approach and primer walking. Data covers the first fourteen 384-well plates received since the beginning of the project.
- Approximately 56% of pooled clones are finished after the first ½ shotgun reads are completed.
- For clones that require a directed strategy after shotgun, primers are generated using Consed[4] and/or Autofinish[5].



Numbers current as of May 01, 2002

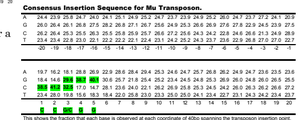
Mu Transposon Target Site Analysis

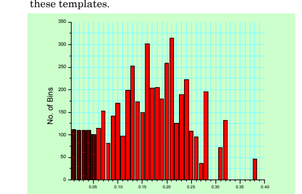




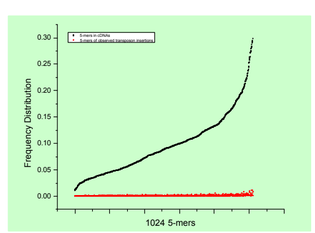
At the site of insertion, there is a 5bp duplication

Consensus Insertion Sequence for Mu Transposon.

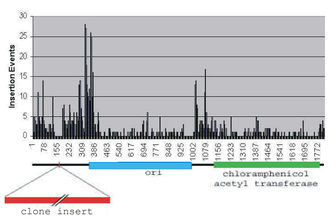




Mu transposon insertion deviates only slightly from random. The insertions of Mu into 1,242 cDNA clones were analyzed using the binomial test and assigned to bins (Materials and Methods). The resulting p-values reflect the likelihood that the observed insertion events were not random. Plotted are the numbers of bins grouped into p-value ranges of 0.01. P-values of greater than 0.05 correspond to bins for which the observed insertion events are likely to be random. P-values of 0.05 or less (indicated by shaded bars) correspond to bins for which the observed insertion events cannot be confidently described as random occurrences.



Comparison of the frequency of 5-mers occuring cDNAs with the frequency of 5-mers utilized in 22,785 transposon insertion events (see Table 2). in the sequences of 1,242



Analysis of transposon insertions in the pOTB7 vector. A set of sequences (5,552) were analyzed and the relative positions of transposon insertions within the vector were mapped. Of the sequences in this set, 22% (1,233) were observed to initiate within the vector. If transposon insertion into the vector was random and all insertion events were recovered, we would expect to observe 46% (2,547) of reads initiating in the vector. However, the observed results match more closely to that expected (16% insertion into vector) assuming zero insertions into the vector chloramphenicol acetyl transferase (*cat*) gene and origin of replication (ori). This is reflected in no tably fewer observed transposon-generated sequence reads initiating from these regions. Transposon insertions into the vector origin of replication are effectively lethal. When chloramphenicol resisistance is included during selection of transposon-containing clones, insertions into the *cat* gene are also effectively lethal. This lethality results in a decreased number of recovered vector insertion events. See Y. Butterfield et al, 2002 , NAR Vol30, #11

Assuming random insertion along entire vector

	Expected	Observed
Insert	3005 (54%)	4319 (78%)
Vector	2547 (46%)	1233 (22%)

Assuming zero insertions in *cat*. and ori.

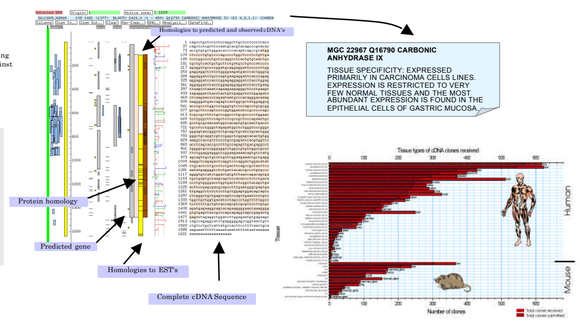
	Expected	Observed
Insert	4676 (84%)	4319 (78%)
Vector	876 (16%)	1233 (22%)

Biological Checks and Annotating

- The coding element and putative function of each finished clone is analysed within ACEDIF.
- Sequence similarity searching is done with each clone against protein and EST databases and alignment of cDNA to genomic sequence.

Biological Analysis

- Full length (poly A tail and start codon)?
- Previously sequenced?
- Frameshift?
- Intronic/genomic sequences?
- No obvious ORF?
- Mitochondrial genomic DNA?
- Duplication?
- Chimeric clone?
- Insertions/deletions?



Summary

Plate #	Received	Days in pipeline	Status	Clone Breakdown Analyzed	Problematic	Remaining
IRAL 6	2000-09-07	584	Done 100% completed	346	38	0
IRAL 8	2000-10-05	556	Done 100% completed	339	45	0
IRAL 9	2000-11-08	522	Done 100% completed	342	42	0
IRAL 13	2000-12-19	481	Done 100% completed	303	81	0
IRAL 18	2001-01-31	438	In Progress 99.7% completed	307	76	1
IRAL 22	2001-03-21	389	In Progress 99.5% completed	283	99	2
IRAL 23	2001-03-21	389	In Progress 99.5% completed	275	107	2
IRAL 29	2001-06-14	304	In Progress 99.2% completed	342	57	3
IRAK 15	2001-06-14	304	In Progress 97.1% completed	320	53	11
IRAL 34	2001-07-24	264	In Progress 98.7% completed	280	99	5
IRAK 38	2001-09-06	220	In Progress 97.7% completed	333	42	9
IRAL 40	2001-10-18	178	In Progress 99.5% completed	303	79	2
IRAK 57	2001-11-27	138	In Progress 85.7% completed	293	36	55
IRAL 42	2002-01-10	94	In Progress 95.3% completed	280	86	18
IRAK 67	2002-02-25	48	In Progress 64.1% completed	170	76	138
IRAK 70	2002-04-12	14	In Progress 10.7% completed	0	41	343
16 plates	**n/a**		**In Progress 90.8% completed**	**4521**	**1057**	**566**

[1]status - covers the progress of the plate in terms of completed clones, not the degree of sequencing done on the plate.
[2]Analyzed - Includes all completed clones which have been analyzed by a finisher and have also been identified as requiring no further sequencing effort to finish. These clones are either submitted or are in the queue to be finished and checked biologically. This category does not include clones that have been marked as problematic but may include clones that eventually may be identified as problematic from the finished sequence.
[3]Problematic - Includes all completed clones that have been marked as problematic and hence require no further sequencing or inspection. These clones are completed and have been or will be placed into a problem clone report.
[4]Remaining - Includes all clones of neither of the above two categories. Either the clones have not yet been through *all* stages of the pipeline yet or further sequencing/finishing is required to finish the clone or identify it as problematic.