

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

A Genomic Approach to Splice Variant Detection, Primer Design, and Identification of Gene Trap Sequence Tags.

Permalink

<https://escholarship.org/uc/item/9jg3h2zt>

Author

Harper, Courtney

Publication Date

2007-09-13

Peer reviewed|Thesis/dissertation

A Genomic Approach to Splice Variant Detection, Primer Design, and Identification of Gene Trap Sequence Tags.

by

Courtney Harper

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

UMI Number: 3274657



UMI Microform 3274657

Copyright 2007 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Copyright (2007)
by
Courtney A. Harper

I dedicate my dissertation to my mother, Rosalind, who is my hero. And to the family and friends who provided such strong support over the years. I love you all.

Acknowledgements

I would like to thank my advisor, Prof. Patricia C. Babbitt, who supported me in my research. The first time I met Patsy, she described the field of bioinformatics as a path strewn with diamonds, and all we had to do was walk ahead and pick them up. I was in the process of choosing a graduate school and had spent the previous weeks in intense and somewhat confrontational interviews with professors who wanted to know why I thought I was good enough for their program. Patsy's optimism and natural inclusiveness was not just refreshing – it was inspirational. I decided upon the spot to join the Program in Biological and Medical Informatics at UCSF, and furthermore to try to join the Babbitt laboratory. I've never for a moment regretted my decision. Patsy has given me invaluable advice, compassionate criticism, and a great deal of inspiration.

Prof. Tom Ferrin has also made great contributions to my progress as a graduate student. He and Patsy welcomed me into the BayGenomics group and were generous enough to let me participate at a low level of involvement when I needed to devote the lions share of my time to other projects while still taking on small duties to gain teamwork and leadership experience. Tom also chaired my orals committee, giving me great advice on how to present and defend my work.

I would also like to thank the other members of my advisory committees, Prof. Andrej Sali, and Prof. Bradford Gibson, and especially Prof. Maria Pallavicini, for all of their advice and help.

I'd like to thank the members of the Babbitt laboratory, past and present. They have been wonderful role models, mentors and friends.

I'd also like to thank the members of the IGTC and BayGenomics groups for allowing me to be a member of such a stellar team.

I would like to thank the lovely people at *BMC Genomics*, who published the manuscript that makes up the main body of Chapter III.

Abstract

The availability of full genome sequences for many organisms has greatly increased the reach of bioinformatics. In my research, I have used a variety of techniques to leverage the information carried in mouse, human, and viral genomes to address a diverse set of challenges.

One challenge was to devise a set of sequences to detect various strains of Human Papillomavirus (HPV). Chapter I describes the method by which I designed probe sequences common to multiple genomes to efficiently isolate HPV DNA from human tissue samples and probe sequences unique to each HPV genome to differentiate between viral strains for the purpose of diagnosing infections.

Chapter II depicts my role in developing the prototype International Gene Trap Consortium web resource, which presents information about embryonic stem cell lines carrying single gene knockouts to the public. Much of this work involved the creation of a new web site and a multi-path process for identification of gene trap sequence tags. Chapter III describes work that developed out of the transition from an mRNA transcript-based sequence tag annotation method to a process that combines transcript matching with localization to the mouse genome. To understand better the localization of gene trap sequence tags to the mouse genome, I compared stand-alone versions of the common genome alignment programs BLAT, SSAHA, and MegaBLAST.

Chapter IV details a method to detect splice variation in different tissues. I developed a process to combine information about splice variants gained by aligning expressed-sequence tags (ESTs) with full-length gene transcripts with microarray analysis to detect splice variants in high-throughput expression data. This method utilized data from pre-existing microarray expression experiments, and so had the potential for large-scale academic and industry use.

Table of Contents

CHAPTER I: INTRODUCTION.....	1
REFERENCES	3
CHAPTER II: PRIMER DESIGN FOR HPV GENOMES.....	4
INTRODUCTION	4
METHODS.....	6
<i>Experimental Overview</i>	6
<i>Bioinformatics Overview</i>	9
<i>Genomes</i>	9
<i>Restriction Enzyme Optimization</i>	9
<i>General Probe Design Parameters</i>	10
<i>Detection Probes</i>	13
<i>Capture Probes</i>	15
CURRENT STATUS.....	19
REFERENCES	21
CHAPTER III: GENE TRAP RESOURCE DEVELOPMENT	22
INTRODUCTION	22
METHODS.....	24
CONCLUSIONS.....	29
REFERENCES	49
CHAPTER IV: COMPARISON OF METHODS FOR GENOMIC LOCALIZATION OF GENE TRAP SEQUENCES	50
MANUSCRIPT	52
ABSTRACT.....	53
BACKGROUND	54
RESULTS AND DISCUSSION	60
<i>Localization to the correct gene</i>	60
<i>Pseudogenes</i>	65
<i>Localization to the correct exon</i>	68
<i>Localization to the correct nucleotide</i>	71
<i>Algorithm speed</i>	75
CONCLUSIONS.....	75
METHODS.....	77
<i>Sequences</i>	77
<i>Computation</i>	78
AUTHOR'S CONTRIBUTIONS.....	83
ACKNOWLEDGMENTS.....	83
REFERENCES	84
CHAPTER V: ANALYSIS OF ALTERNATIVE SPLICING UTILIZING MICROARRAY EXPERIMENTS	86
INTRODUCTION	86
BACKGROUND	86
METHODS.....	92
<i>Expression Data Set</i>	94
<i>Association with Splice Variants</i>	95
<i>Splice Variant Detection Method</i>	97
RESULTS.....	99
CONCLUSIONS.....	104
REFERENCES	108

List of Tables

Table 2.1	The number of probes generated for each risk-associated HPV genome.	16
Table 4.1.	Computation times in seconds for each localization program.....	63
Table 4.2.	Parameters for each localization program used.....	82
Table 5.1.	A breakdown of experimental results by Affymetrix Gene Chip.....	98

List of Figures

Figure 2.1. Overview of the HPV detection method.....	8
Figure 2.2. Overview of the capture probe creation process.	18
Figure 3.1. The design for data flow through the IGTC web site.....	26
Figure 3.2. Graphic of the proposed IGTC sequence tag annotation web page.	27
Figure 3.3. Graphic of the proposed IGTC gene annotation web page.....	28
Figure 3.4. The IGTC Overview Tutorial.....	31
Figure 3.5. The IGTC Search Tutorial.	35
Additional Figure 4.1. An example of errors associated with low signal strength.	56
Figure 4.1. Recall and precision for each localization algorithm.	61
Figure 4.2. An example of localization to a pseudogene.....	66
Figure 4.3. A representative alignment of a full-length gene and a sequence tag.	69
Figure 4.4. A summary of the alignments by each program to the edges of exons.	73
Figure 5.1. The three types of alternative splicing analyzed in this research.	88
Figure 5.2. Overview of the microarray experimental process.....	90
Figure 5.3. Graphical representation of probes aligned to two transcripts.....	93
Figure 5.4. Intragenic hybridization value differences.....	100
Figure 5.5. Distribution of hybridization values.	102
Figure 5.6. Mapping of 17,520 probe sets by average hybridization value.....	103
Figure 5.7. The number of probes in conserved and variant regions.	105
Figure 5.8. Cumulative distribution of hybridization value differences.	106

Introduction

The availability of full genome sequences for many organisms has opened new avenues of research in many fields [1-3]. Researchers seeking to understand the complexities that underlie the process by which information encoded in genes transitions to functioning proteins can use the intronic and flanking sequences that surround coding regions to search for splice sites, transcription factor binding sites, or other elements that control transcription and translation [4, 5]. Alignments between RNA transcripts and their source genome can be used to aid in gene annotation and the prediction of alternative splicing [6]. Alignment of different genomes can illustrate conserved elements involved in important gene functions or indicate sequences that can be used to differentiate between similar organisms [7]. These are but a few examples of the wide range of practices making use of genomic sequence to address biological questions. In my research, I have used some of these techniques to address a varied set of biological questions.

One such biological question that I used genomic information to tackle involved detection of the Human Papillomavirus (HPV) for the purpose of diagnosing infections in human tissue. There are 105 HPV genomes that have been sequenced to date, of which 23 HPV genomes are known to be associated with genital warts or cervical cancer [8]. Although there is currently a vaccine for four types of HPV [9], detection methods for all HPV types will continue to be useful for strains not covered by the vaccine and for populations that already have HPV. Chapter I describes the method by which I devised a set of sequences common to multiple genomes to efficiently isolate HPV DNA from

human tissue samples, as well as probe sequences unique to each HPV genome to differentiate between viral strains.

My work with the gene trapping group BayGenomics [10] allowed me to use genomic analysis to aid gene annotation. Chapter II depicts my role in assisting the BayGenomics group and helping to develop the prototype International Gene Trap Consortium web resource [11]. Part of my role was to design a genomic localization based protocol for identification of gene trap sequence tags. To understand better the localization of gene trap sequence tags to the mouse genome, I compared stand-alone versions of the genome alignment programs in use at three major genome browser web sites. Chapter III consists of a manuscript summarizing this research and presenting the pitfalls of aligning short, poor quality sequence such as gene trap sequence tags to the mouse genome.

An interest in alternative splicing led me to combine two whole-genome approaches, microarray expression analysis and multiple alignment of expressed-sequence tags (ESTs) with mRNA sequences, to better predict splice variation. Chapter IV details a method I developed to detect splice variants in pre-existing microarray data. This method takes advantage of the fact that microarrays contain many probes for a single gene, and attempts to use differences in fluorescent signal across the length of a gene to detect the presence of a known splice variant. The known splice variants were generated by alignment of ESTs and mRNAs in the Unigene database [12]. This method had potential for widespread use, given the large amount of microarray expression data generated for other purposes that could be secondarily mined for splice variant information.

References

1. Debes JD, Urrutia R: **Bioinformatics tools to understand human diseases.** *Surgery* 2004, **135**(6):579-585.
2. Desany B, Zhang Z: **Bioinformatics and cancer target discovery.** *Drug Discov Today* 2004, **9**(18):795-802.
3. Ivakhno S: **From functional genomics to systems biology.** *Febs J* 2007, **274**(10):2439-2448.
4. Chua G, Robinson MD, Morris Q, Hughes TR: **Transcriptional networks: reverse-engineering gene regulation on a global scale.** *Curr Opin Microbiol* 2004, **7**(6):638-646.
5. Zhou D, Yang R: **Global analysis of gene transcription regulation in prokaryotes.** *Cell Mol Life Sci* 2006, **63**(19-20):2260-2290.
6. Lee C, Wang Q: **Bioinformatics analysis of alternative splicing.** *Brief Bioinform* 2005, **6**(1):23-33.
7. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: **Towards fully automated structure-based function prediction in structural genomics: a case study.** *J Mol Biol* 2007, **367**(5):1511-1522.
8. Bernard HU, Calleja-Macias IE, Dunn ST: **Genome variation of human papillomavirus types: phylogenetic and medical implications.** *Int J Cancer* 2006, **118**(5):1071-1076.
9. Siddiqui MA, Perry CM: **Human papillomavirus quadrivalent (types 6, 11, 16, 18) recombinant vaccine (Gardasil).** *Drugs* 2006, **66**(9):1263-1271; discussion 1272-1263.
10. Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L *et al*: **BayGenomics: a resource of insertional mutations in mouse embryonic stem cells.** *Nucleic Acids Res* 2003, **31**(1):278-281.
11. Nord AS, Chang PJ, Conklin BR, Cox AV, Harper CA, Hicks GG, Huang CC, Johns SJ, Kawamoto M, Liu S *et al*: **The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse.** *Nucleic Acids Res* 2006, **34**(Database issue):D642-648.
12. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** *J Mol Med* 1997, **75**(10):694-698.

Primer Design for HPV Genomes

Introduction

Human papillomavirus (HPV) is a member of the Papillomaviridae family of DNA viruses. Papillomaviridae viruses live within a host, infecting the epithelial cells, and are shed through mucosal membranes. Many animal species are known to carry different strains of HPV, although each strain is infectious only within a single species. There are at least 105 distinct genomes identified that infect humans, numbered in the order they were discovered, with no two genomes sharing greater than 90% sequence identity in coding regions [1].

HPV sequences mutate rarely, and so the sequences of an HPV genome can be expected to remain stable throughout the duration of an infection, with the exception of breakage in the HPV genome surrounding the site at which the circular genome is cleaved prior to integration into the human genome. This stability of HPV genomes is inferred from a phylogenetic study which clustered HPV genomes and their homologs in various animals by the sequence of the L1 gene. The presence of both HPV and animal sequences in the same clusters indicates that individual HPV genomes share more sequence similarity with the homologous genomes in other species than with other HPV genomes [2].

HPV infection is the most common sexually transmitted disease in the United States. There are approximately 5.5 million new cases each year [3], and it is estimated that 50% - 80% of women will acquire one or more types of HPV during their lifetime [4,

5]. HPV is shed through mucosal membranes, and can be transmitted through oral and anal sex as well as vaginal intercourse. Condom use has not proven to be effective in preventing transmission of the HPV virus [6], although it may offer partial protection with vigilant use [7].

Most HPV infections are asymptomatic and carry no long-term risk, but some infections lead to genital warts, cervical cancer, and more rarely other cancers, such as penile or anal cancer. HPV causes 99% of cervical cancer, the second most common cancer in women [8]. The virus causes cancer by expressing genes that interact with cell-cycle controls to allow viral replication and release. These protein interactions can lead to over-proliferation of cells and prevention of apoptosis.

Cervical cancer can be detected in the early stages by Papanicolaou (Pap) testing, where epithelial cells gathered from a patient's cervix are visually observed for evidence of dysplasia or cell abnormality. The Pap smear can detect precancerous lesions before they develop into cancer, but this procedure is only ~80% accurate for dysplasia diagnosis. Further steps are required in response to an abnormal Pap smear, including additional Pap smears to confirm the initial dysplasia result and extensive procedures to determine if a high risk or a low risk type is present. This is a costly and lengthy process, but treatment for the early stages of cervical cancer is highly effective in preventing mortality, so Pap testing is considered a necessary part of basic women's health care.

A collaborator at the University of California Santa Barbara, Dr. Norbert Reich, has proposed methods of detecting and typing HPV infections in human tissue. The technology developed in his laboratory relies on probes that will hybridize with and amplify the signal of HPV DNA sequences unique to a single HPV genome, allowing the

direct identification of any known HPV type from a clinical sample. This technology has seeded a biotechnology start-up company called Tamarisc Diagnostics, Inc. Their goal is to develop an assay for use by physicians that is inexpensive compared to Pap testing and fast enough to be completed in a single patient appointment. In order to accomplish this goal, Tamarisc Diagnostics, Inc. required two sets of probes to be designed by a bioinformaticist. In collaboration with the Reich Lab and Tamarisc Diagnostics, Inc., I developed these probes. The first set would be used to capture HPV sequence in a clinical sample through DNA-RNA hybridization. The second set would be used to determine which type of HPV sequence is present in a clinical sample.

Methods

Experimental Overview

A general overview of the HPV detection method developed by Tamarisc Diagnostics, Inc. is shown in Figure 1. Initially, a swab sample is taken from a patient, and the sample is transferred to an apparatus containing all of the necessary reagents to perform the detection assay. Cells contained in the sample are lysed and the DNA they contain is cut with a restriction enzyme to yield a mixture of fragments of human genomic DNA, HPV sequence that has integrated into the human genome, and circular HPV DNA that remains separate from the human genome. Linear and circular HPV

DNA are separated from human genomic DNA by hybridization with probes specific to HPV sequences. After this purification step, probes specific for individual HPV genomes are added. These RNA probes hybridize to complementary DNA, displacing the DNA already present and disrupting the double helix. An enzyme is then added that detects this disruption and activates a fluorescent marker in response to the RNA-DNA duplex. This fluorescence is then measured, indicating the presence or absence of a specific strain of HPV in the patient sample.

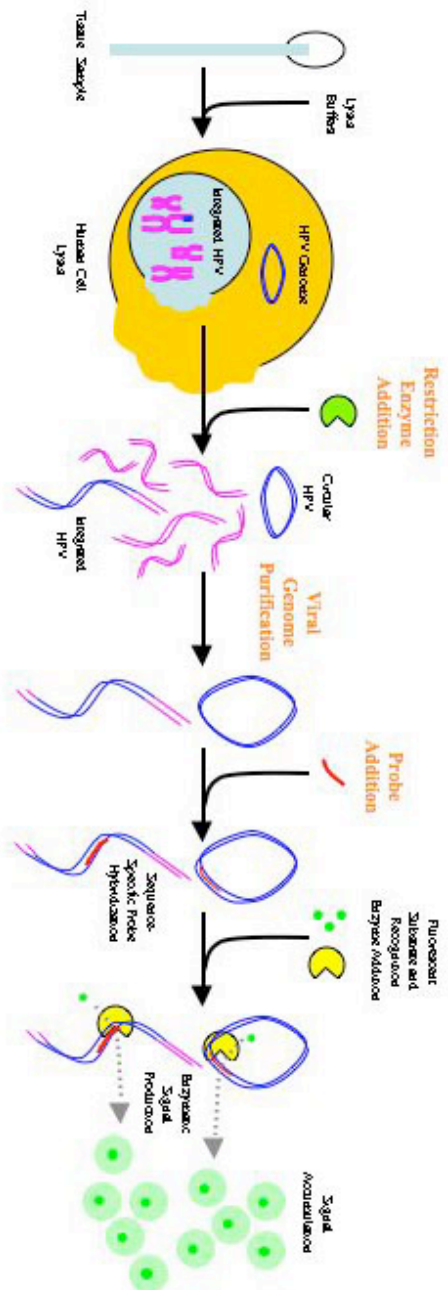


Figure 1. (Adapted from a diagram created by Dr. Norbert Reich and Dr. August Estabrook.) An overview of the HPV detection method, with steps involving the work described in this chapter listed in orange.

Bioinformatics Overview

The bioinformatics effort for Tamarisc Diagnostics, Inc. was centered on probe design. The aim was to find all known HPV genomes and devise a minimal set of capture probes to purify HPV genomic sequence from clinical samples and one or a small set of detection probes for each HPV genome that will uniquely identify a sample infected with that strain of HPV. As HPV genomes are relatively stable, it is not expected that primer sequences will need to change very often.

Genomes

All HPV genomes were collected from the HPV databases at Los Alamos National Laboratory (<http://hpv-web.lanl.gov>) and the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov>). These sequences represent HPV genomes 1 through 106, excluding 46, 64, 78, 79, 85, 88, 98, 99, 104 and 105, for which full genomic sequences were not available. The probe sequences generated represent 23 HPV genomes that represent risk to a patient; 1 genome associated with both genital warts and cervical cancer, 5 genomes associated with genital warts, and 17 genomes associated with cervical cancer, including HPV 16 and HPV 18.

Restriction Enzyme Optimization

Before probe design began, a restriction enzyme had to be chosen to remove the human genome from the cellular extract. This step is necessary to minimize the chance that signal from incomplete or inexact hybridization between the probes and the human genome would swamp real signal due to the tremendous size of the human genome, which contains approximately 3 billion base pairs, compared with HPV genomes, which average 8000 base pairs. This step also allows the probe design steps that follow to avoid selection of probes containing cuts sites for the chosen restriction enzyme. The first step in isolating HPV sequences is to cut the human genome into smaller pieces with a restriction enzyme while leaving HPV DNA relatively intact. Sixty-two commercially available enzymes were tested for their cutting frequency in both human and HPV genomes. It was determined that although BstZI and EagI cut only three HPV genomes once and the remaining HPV genomes not at all, they do not cut the human genome frequently and are methylation sensitive, making them inappropriate for experimental conditions. Instead, BglII was chosen because it is methylation-insensitive and cuts only eight HPV genomes while cutting the human genome frequently.

General Probe Design Parameters

The design of the capture probes and the detection probes contained many of the same steps, since both involve single-strand RNA probes that will form hetero-duplexes with viral DNA. These factors that required consideration included the length of the probe, the melting temperature (T_m) of the probe, the uniqueness of the probe, any potential to form a stable secondary structure or self-hybridize, whether the chosen

restriction enzyme would cut at the same site at which the probe would hybridize, and whether the probe is likely to hybridize to sequences other than the target sequence.

The probes must be sufficiently long that the cumulative stabilization energy from each hybridized base pair is sufficient to maintain the hetero-duplex long enough for RNaseH to degrade the RNA portion of the probe. Hetero-duplex formation occurs as an equilibrium reaction, with nucleotides binding to and dissociating from each other at rates that are dependent on temperature, concentration of the sequences in solution, and whether neighboring nucleotides are already hybridized. Lower temperatures and higher concentrations of probe and target sequences generally favor hetero-duplex formation. Once hybridization has begun, further binding is favored, as the effective concentration of unhybridized nucleotides is greatly increased by being tethered to hybridized sequence. This can pose a problem in terms of probe specificity. With short probes, any hetero-duplexes formed between a probe and a sequence that does not exactly match the target sequence will dissociate too rapidly to be degraded by RNaseH, simply because there are not enough matching nucleotides to create a sufficient binding strength. However, long probes may allow hetero-duplexes containing one or more mismatches to remain hybridized long enough for the probe to be degraded. Initially, test probes between 20 and 30 nucleotides in length were created, but probes of differing lengths can rapidly generated should it become necessary.

In order to minimize signal produced by hetero-duplexes containing mismatches, the reactions will be performed at a temperature that is slightly lower than the melting temperature (T_m) of each probe, which is the temperature at which dissociation and hybridization of the probe and the target sequence are equally favored. An acceptable

T_m range for probes with 20 to 30 nucleotides was set from 60°C - 90°C, a temperature range that is compatible with thermostable RNaseH activity [9]. The goal is then to maximize the difference between the T_m of the probe-target hetero-duplex and the T_m of any hetero-duplexes formed between the probe and non-target sequence. Requiring a short probe length, so that mismatches have a proportionally larger affect on binding strength, is helpful. Even more effective is limiting the number of consecutive matching nucleotides between a probe sequence and a non-target sequence. In this case, non-target sequence would be any DNA purified from a patient sample, most likely other HPV strains or viral genomes. A maximum of 11 consecutive matches was allowed between a probe and genomic sequence. This threshold is short enough to be easily differentiable from a correct hybridization of 20 or more nucleotides by the difference in T_m .

Additionally, it is beneficial to choose probe sequences that will not self-hybridize or form stable secondary structure. It is nearly impossible to design probes that will not form small hairpins, but large regions of complementarity should be avoided. A threshold of 11 nucleotides was set as the maximum consecutive complimentary nucleotides allowed between two copies of a probe, or within a single probe. More complex secondary structure is more difficult to predict, especially given that non-nucleotide compounds are bound to either end of the probe with the HET technology. Predicting probes that will form secondary structure beyond hairpins or helices was not a priority, since such probe sequences will not cause a false-positive signal, although they will not participate in hetero-duplexes while they are non-linear, thus somewhat dampening a true signal.

Finally, all probes that contain a cut site for the BglIII restriction enzyme must be excluded, as that enzyme will be used in a step prior to the probe hybridization step.

Detection Probes

Several probe-design programs were evaluated to determine if an existing program would be able to meet all of the probe design parameters detailed above. No single program has all the functionality required, but the publicly available program OligoPicker [10] does have many useful features. OligoPicker allows a user to create probes of a given length, screen for matches to a small set of non-target sequences, tests for self-complementarity, and calculates a predicted T_m for each probe. However, the OligoPicker program does not meet several requirements, including identification of many potential non-target matches, elimination of repetitive nucleotides from probes, and setting of minimal T_m . Most perplexingly, it only generates a maximum of 5 probes per run. To supplement its capabilities, OligoPicker was run from within a program written using the Python programming language to exhaustively search each risk-associated HPV genome for probes that met length, T_m , maximal matching region, exclusion of BglIII cut sites, and self-hybridization criteria. Table 1 shows the number of probes for each HPV genome that met these initial selection criteria.

Although the majority of HPV DNA takes the form of a circular, double-stranded genome, HPV sequences are known to integrate into the human genome. While the Tamarisc Diagnostics, Inc. technology was developed for the identification of discrete HPV sequences, steps have been taken to generate probes that would be equally useful

for the detection of integrated sequences. The human genome contains approximately 3 billion base pairs, vastly increasing the likelihood that a 20-30 nucleotide RNA probe will hybridize at least partially to some region of the human genome, creating the potential for a false positive reading. In fact, all probes partially or fully match at least one chromosomal sequence. Additionally, other genomes, such as those associated with sexually transmitted diseases (STDs), may be present in a tissue sample and contain sequence that will hybridize with a probe.

Potential hetero-duplexes between probe and non-target sequences were determined by using the BLAST program [11] to search for partially or fully complementary sequence in a screening set of the human genome and sequences from Chlamydia, herpes, adenovirus, trichomonas, gonorrhea, HIV, and any HPV genome other than the target of a particular probe. BLAST was chosen because it is capable, with the right parameter settings, of rapidly detecting very distant matches, some of which boast as little as 60% sequence identity with a search probe. In order to detect short partial matches, the BLAST gap opening and gap extension penalties were set to -1, the mismatch penalty was set to -1, and the expect value was set to 10,000. In order to deal with the tremendous number of matches returned by such a lenient search, a Python script was written to automatically parse the results for alignments with 11 or more matching nucleotides, which could be sufficient to result in hybridization between a probe and non-target sequence. All probes had matches to non-target sequences that exceeded 11 nucleotides over the length of the probe.

In order to determine which probes are least likely to create a fluorescent signal in the presence of non-target DNA, the predicted T_m of each probe-non-target hetero-

duplex was compared to the predicted T_m of the probe-target hetero-duplex. The results were ranked by T_m difference, and a threshold of 10 °C minimum difference was chosen to balance the need for a large temperature range in which to experimentally optimize detection and the need to retain enough acceptable probe sequences to provide replacements for probes that fail. The number of probes for each genome that meet this criterion are listed in Table 1.

As Table 1 demonstrates, after rigorous selection criteria for single-strand RNA probes were met, a sizeable number of probe options remained for each HPV genome. The process used to generate these probes has been largely automated, so the addition of new HPV genomes or screening sequences would not present a problem in terms of probe-design. Additionally, all data generated during the probe-design process have been retained, so altering a particular selection threshold does not necessitate running the programs anew.

Capture Probes

To purify HPV genomic sequence from clinical samples, a set of probes was needed that would hybridize with all HPV genomes while avoiding hybridization with

HPV genome number and accession	Risk type	# probes meeting initial selection criteria	# probes with Tm 10°C higher than closest mismatch
HPV-6b 9626053	High, Low	1410	78
HPV-11 M14119	Low	1501	99
HPV-16 9627100	High	1446	*7
HPV-18 9626069	High	1407	86
HPV-31 J04353	High	1535	155
HPV-33 M12732	High	1208	153
HPV-35 X74477	High	1446	123
HPV-39 M62849	High	1272	130
HPV-40 X74478	Low	1591	228
HPV-42 M73236	Low	1711	202
HPV-43 40804474	Low	1396	159
HPV-44 U31788	Low	450	39
HPV-45 X74479	High	1467	73
HPV-51 M62877	High	1261	148
HPV-52 X74481	High	1591	200
HPV-56 X74483	High	1066	92
HPV-58 D90400	High	1323	85
HPV-59 X77858	High	1737	203
HPV-66 U31794	High	1167	78
HPV-68a 71726685	High	1258	70
HPV-70 U21941	High	1344	133
HPV-73 X94165	High	1146	105
HPV-82 6970427	High	1301	64
Total number of probes		31034	2703

Table 2.1 The number of probes generated for each risk-associated HPV genome. The middle column contains the number of probes for a given HPV genome that met the initial selection criteria: length of 20-30 nucleotides, predicted T_m of 60°C - 90°C, and no subsequence greater than 11 nucleotides matching to another HPV genome, viral sequence, self-probe copy, BglII cut site, or internal match. The rightmost column contains the number of probes that, in addition to meeting selection criteria, do not match any sequence in the human genome closely enough to have a predicted T_m of a probe-non-target hetero-duplex within 10°C of the predicted T_m of the probe-target hetero-duplex. * If the T_m difference threshold is lowered to 8 °C for HPV-16, 34 probes meet this criterion.

human genomic or STD sequences. A minimal set was desired in order to reduce the occurrence of non-target hybridization that could result in the purification of non-HPV sequences. At this point, only capture probes for high- and low-risk HPV genomes have been created, although the process can be repeated with more genomes should that be desired. An overview of the capture probe design process is shown in Figure 2.

To determine which probes would be useful in such a probe set, every possible probe of a given length was generated from each HPV genome, resulting in probes tiled over the length of the genome. These probes were filtered to remove duplicate sequences within a genome, probes containing strings of five or more identical consecutive nucleotides, probes predicted to form hairpin structures, probes containing BglIII recognition sequences, and probes with a predicted T_m of less than 60°C. The remaining probes were aligned with the same set human genomic and STD sequences used to create the detection probes. This comparison was performed with BLAST using gap opening and extension penalties of -1, a mismatch penalty of -1, and an expect value of 10,000. Probes with fewer than five contiguous mismatches or 7 mismatches over the length of the probe were removed from consideration.

To find a minimum set of capture probes, all probes remaining after filtering for general probe parameters and similarity to human and STD sequences were tested for presence in multiple genomes. The probes were rank-ordered by the number of HPV genomes containing the probe sequence. The top 20 probes were used to seed minimal probe sets. After the first sequence of each minimal

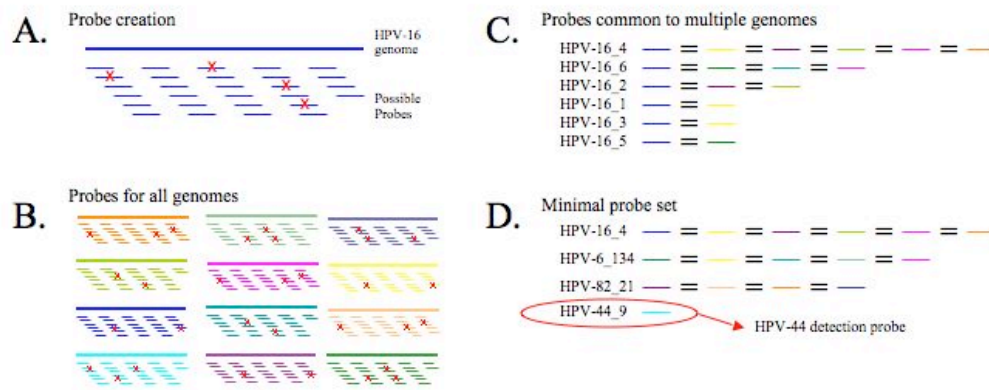


Figure 2.2. Overview of the capture probe creation process. A) Every sequence of a given length was extracted from an HPV genome, excluding duplicate probes and probes that did not meet required parameters. B) This process was repeated for all HPV genomes, resulting in a list of probes for each HPV strain. C) Each probe was tested for presence in multiple genomes. D) Probes present in the most genomes were used to seed minimal probe sets. When a single genome was missing, a detection probe was used.

set was chosen, the next sequence was determined by selecting from all remaining probes the sequence that matched the greatest number of genomes previously unmatched. Each minimal set was completed when all genomes were matched by at least one probe. This process does not guarantee the smallest minimum set possible, but is far more efficient than a full optimization.

Every minimal set of capture probes contained at least one probe that only matched a single HPV genome not previously matched by another probe in that set. In these cases, a detection probe was substituted for the last probe chosen for the minimal set. Using a detection probe to match the missing genome is preferable to using another probe that matches multiple genomes because it allows for optimization of experimental protocols around one sequence instead of two. It also provides some redundancy in the detection method, ensuring that for that HPV genome, any signal detected at the capture stage would be reflected at the detection stage.

Three of the 20 minimum sets contained ten probes that could collectively hybridize with the 23 risk-associated HPV genomes. The remaining minimum sets contained more sequences. Of the three smallest minimum sets, the one with the greatest redundancy in matches was chosen as the candidate capture probe set.

Current Status

Tamarisc Diagnostics, Inc. is currently using a set of one capture probe and two detection probes for the HPV-16 and HPV-18 genomes to determine the appropriate

experimental conditions. Once the process has been optimized sufficiently, sequences from further HPV genomes will be tested and those that work may be used for the development of an HPV detection kit.

References

1. Munger K, Howley PM: **Human papillomavirus immortalization and transformation functions.** *Virus Res* 2002, **89**(2):213-228.
2. Howley PM LD: **Papillomaviruses and their replication.** Philadelphia: Lippincott Williams & Wilkins; 2001.
3. Cates W, Jr.: **Estimates of the incidence and prevalence of sexually transmitted diseases in the United States.** **American Social Health Association Panel.** *Sex Transm Dis* 1999, **26**(4 Suppl):S2-7.
4. Koutsky LA, Galloway DA, Holmes KK: **Epidemiology of genital human papillomavirus infection.** *Epidemiol Rev* 1988, **10**:122-163.
5. Myers ER, McCrory DC, Nanda K, Bastian L, Matchar DB: **Mathematical model for the natural history of human papillomavirus infection and cervical carcinogenesis.** *Am J Epidemiol* 2000, **151**(12):1158-1171.
6. Holmes KK, Levine R, Weaver M: **Effectiveness of condoms in preventing sexually transmitted infections.** *Bull World Health Organ* 2004, **82**(6):454-461.
7. Winer RL, Hughes JP, Feng Q, O'Reilly S, Kiviat NB, Holmes KK, Koutsky LA: **Condom use and the risk of genital human papillomavirus infection in young women.** *N Engl J Med* 2006, **354**(25):2645-2654.
8. Baseman JG, Koutsky LA: **The epidemiology of human papillomavirus infections.** *J Clin Virol* 2005, **32 Suppl 1**:S16-24.
9. Haruki M, Nogawa T, Hirano N, Chon H, Tsunaka Y, Morikawa M, Kanaya S: **Efficient cleavage of RNA at high temperatures by a thermostable DNA-linked ribonuclease H.** *Protein Eng* 2000, **13**(12):881-886.
10. Wang X, Seed B: **Selection of oligonucleotide probes for protein coding sequences.** *Bioinformatics* 2003, **19**(7):796-802.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.

Gene trap Resource Development

Introduction

Studies of the inactivation of genes in a model organism such as mouse can result in great insights into gene function and show the involvement of genes in common diseases. High-throughput, untargeted interruption of genes can be used to create an invaluable resource for scientists without requiring the extensive background knowledge of a gene and considerable input of time needed for targeted gene inactivation. Gene trapping is a method of randomly generating embryonic stem cells with a single interrupted gene. In this method, a gene trap vector construct is inserted into an intronic or coding region of the genome. The vector constructs contains a reporter tag that can be used to identify cell lines where the vector has inserted into a genic region, preventing that gene from being normally transcribed and translated into a functional protein. Since gene trapping is a random process, it allows for the disruption of novel as well as known genes. Gene trapping has proven to be a very reliable process, creating knockouts with phenotypes equivalent to targeted knockouts of the gene in 91% of test cases [1], and taking up to two orders of magnitude less time than a targeted knockout.

BayGenomics is a large undertaking to create and analyze thousands of gene trapping events, and to provide information and cell lines to the public [2]. BayGenomics is part of the Program in Genomics Applications (PGA) funded by the National Heart, Lung, and Blood Institute [3]. BayGenomics member laboratories are currently creating knockout stem cells and mice, determining the phenotype caused by interrupted genes, and studying the expression of mouse genes during development using *in situ*

hybridization. I participated in the bioinformatics component of BayGenomics, which is charged with bioinformatic analysis and interpretation of many of these data, as well as presenting this information through an online interface.

My primary role with the BayGenomics bioinformatics group was as a mediator between end users and the computer scientists who created the BayGenomics web resource. I attended bi-weekly meetings and gave my opinion on various issues having to do with the gene annotation process and usability of the web site. In addition to this, I took on projects such as creating a glossary of terms used on the BayGenomics web site, manually annotating unidentified gene trap sequence tags, or tracking down causes of misannotation. I took an active role in the teaching component, serving as an assistant for several of the bioinformatics training courses offered by BayGenomics, and adapting Prof. Patsy Babbitt's lecture on sequence analysis for a PGA conference. Some of the results of these activities are presented in the following manuscript:

Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L, Chuang PT, Young SG, Skarnes WC, Babbitt PC, Ferrin TE: **BayGenomics: a resource of insertional mutations in mouse embryonic stem cells**. *Nucleic Acids Res* 2003, **31**(1):278-281.

When the time came to expand BayGenomics from a national resource to an international consortium of laboratories performing gene trapping in mice, I was offered a rare opportunity for a graduate student: the chance to lead a group project. I worked with Alex Nord, who was the principal interactor between the different gene trapping laboratories, to design a prototype web resource for the newly created International Gene Trap Consortium (IGTC) [4]. We were tasked with overseeing the initial development of the IGTC web site www.genetrap.org in preparation for a conference being organized by

Professors Ferrin and Babbitt to gather international leaders in gene trap research. Mr. Nord's and my objective was to have a functional database and web interface for gene trap data available by the time of the conference that would incorporate data from the laboratories of the attending gene trap researchers.

Methods

A number of objectives had to be achieved in order to develop a web resource representative of the direction we wanted to head with the IGTC. We planned to use the BayGenomics database and web site as a template to design the IGTC resource, with significant alterations. Some of these changes included modification of the data collection protocols and database setup, the website content and design, and the sequence tag identification protocol.

Members of the BayGenomics bioinformatics group were recruited for the IGTC development group to help design and execute these necessary changes. In addition to Mr. Nord and I, Prof. Conrad Huang, Michiko Kawamoto and Doug Stryke worked to create the prototype IGTC web resource, with Susan Johns working in parallel to create sequence tag alignment representations that would be incorporated into the IGTC gene and cell line web pages. Doug Stryke was in charge of creating the IGTC database and populating it with data from BayGenomics and other resources. Conrad Huang wrote programs to ensure that the IGTC could process new data and update annotations on a regular basis. Michiko Kawamoto took responsibility for much of the HTML coding of the website, including providing database access tools and designing the IGTC logo.

During a series of meetings I organized, the IGTC development group combined their detailed knowledge of the resources available at US organizations such as the National Center for Biotechnology Information (NCBI) [5] and the Mouse Genome Database [6] with Mr. Nord's experience with Ensembl to devise a data-flow pathway from acquisition of raw sequence data to presentation of annotated cell line information to end users. The mockup of his initial design is shown in Figure 1.

I took responsibility for designing the web pages associated with gene and sequence tag annotations generated by the IGTC. Initial mockups of these pages are shown in Figure 2 and Figure 3. From these designs, Michiko developed the HTML code for the majority of the web pages available at the IGTC web site. The purpose of these web pages is to present scientists with a central source of information about knockouts of their gene of interest. As well as providing useful information, these annotation elements allow users to search for a gene or cell line using a wide range of peripheral information. Additionally, Mr. Nord and I created a set of web pages to illustrate the gene trapping process and to explain how to use the IGTC resource. Current versions of these tutorials are available at www.genetrap.org/tutorials, and versions from June 2007 are shown in Figures 4 and 5.

The chief reason for altering the sequence tag annotation protocol was to incorporate the best parts of the identification methods used by the different members of the IGTC. Two sequence tag identification protocols were considered especially useful: AutoIdent, a gene transcript-based protocol developed for BayGenomics primarily by Prof. Huang, and MapTag [7], a genomic localization protocol developed for Ensembl [8]. AutoIdent had been in use for three years at the time, and had been thoroughly

IGTC Website/Database Flow Chart

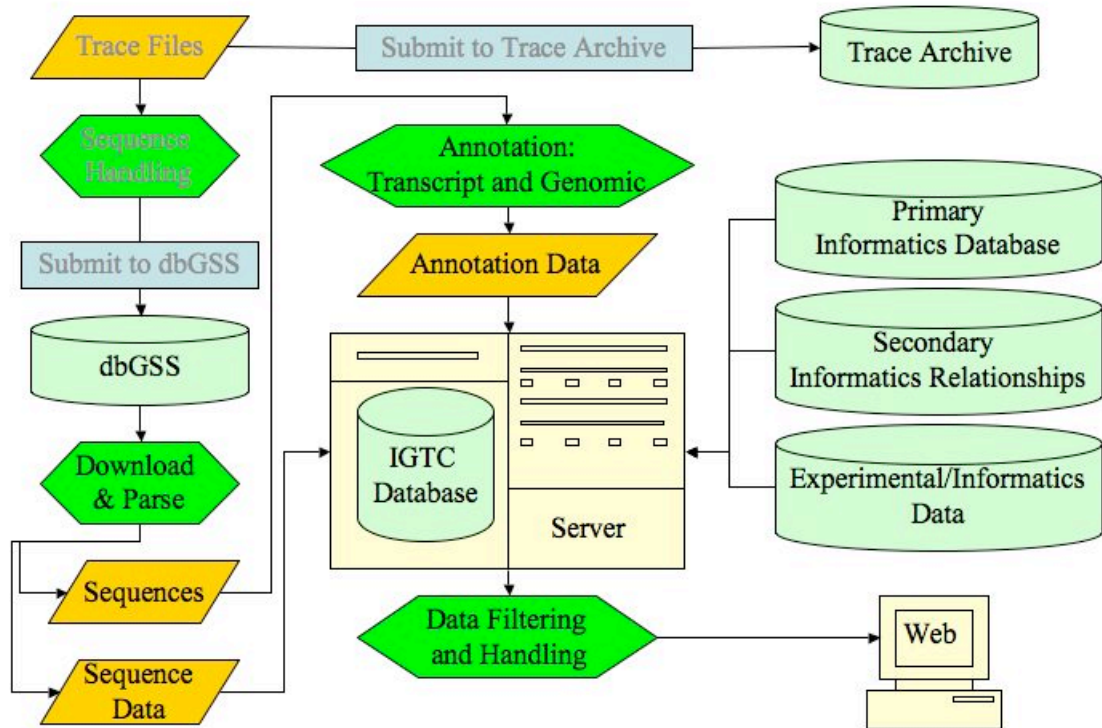


Figure 3.1. Alex Nord’s design for data flow through the IGTC web site. Elements outside of the IGTC pipeline are shown in grey text, whereas constituents of the IGTC resource are shown in black text.

Sequence Tag Report

Primary Annotation Data

Sequence Tag ID
Putative ID: Gene Symbol and Gene Descriptor
Annotation Confidence: Confirmed | Unconfirmed
Explanation:

Genomic Annotation: Map Coordinates
Ensembl ID / Entrez:Gene ID
Match Class
Score

Transcript Annotation:
Transcript Accession ID
Match Class
Score

[Detailed Report](#)

External Links

Annotation Source: Un/Confirmed > Genomic | Transcript
Other Accessions Gene Ontology
Orthologues Protein
Phenotype GenMAPP
Expression Data
Etc.

Sequence Tag

Source
Vector
Quality
Sequence
Trace
History: (new ID, category change, etc.)

Figure 3.2. Graphic representation of the proposed IGTC sequence tag annotation web page.

Gene Report

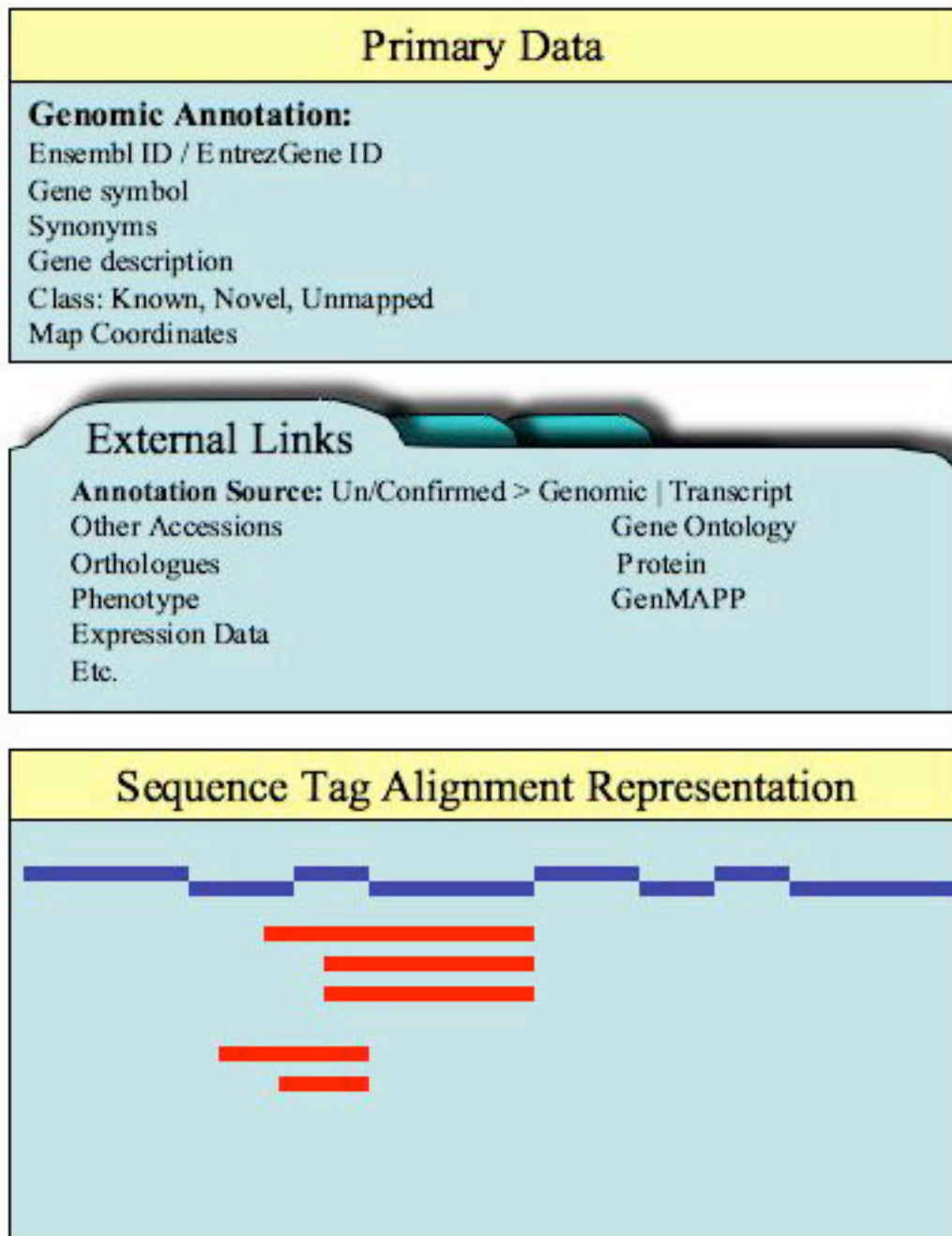


Figure 3.3. Graphic representation of the proposed IGTC gene annotation web page.

validated [2]. MapTag was developed more recently, and was undergoing significant changes that would not result in a stable build in the near future. Therefore, we agreed that the most sensible method of merging the protocols was to use AutoIdent in parallel with a genomic localization protocol of our own design, and to reconcile the results of each identification path. I performed a great deal of research to determine which localization program would be best suited for localizing the IGTC sequence tags to the mouse genome. This research is detailed in Chapter 4. I concluded that the best solution was to use the Blast-like Alignment Tool (BLAT) available at UC Santa Cruz [9] to perform the genomic localization of sequence tags, and wrote a script to do so that was run in parallel to AutoIdent. As matches to gene transcripts and genomic localization are orthogonal data, although both based on sequence alignment, using both annotation protocols adds a level of confidence to sequence tags yielding equivalent results with both programs. Doug Stryke developed a reconciliation process based upon finding correspondence between the genomic localization of sequence tags and the genomic localizations of any matching full-length transcripts.

Conclusions

The prototype web resource was successfully presented at the inaugural IGTC conference on April 15, 2005. The initial version of the database contained over 33,000 cell lines collected from researchers in six countries. Most of the steps involved in the data collection protocols were successfully completed by the time of the conference, and those that were not had pre-computed data inserted in their place. The sequence tag

identification protocol and website content were discussed, with minor changes made where necessary. After incorporation of these changes, Alex Nord drafted a manuscript detailing the IGTC resource:

Nord AS, Chang PJ, Conklin BR, Cox AV, Harper CA, Hicks GG, Huang CC, Johns SJ, Kawamoto M, Liu S, Meng EC, Morris JH, Rossant J, Ruiz P, Skarnes WC, Soriano P, Stanford WL, Stryke D, von Melchner H, Wurst W, Yamamura K, Young SG, Babbitt PC, Ferrin TE: **The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse.** *Nucleic Acids Res* 2006, **34**(Database issue):D642-648.

The current version of the IGTC web resource presents more than double the original number of cell lines and contains far more information about the trapped genes available, but retains much of the structure developed for the prototype used at the conference.

Figure 3.4. The following 4 pages contain the contents of the IGTC Overview Tutorial, which displays information about gene trapping. Below is the first page of the tutorial.

International Gene Trap Consortium

INFORMATION
DATA ACCESS
TUTORIALS
REQUEST ES CELL LINES

Overview

Gene Trap Overview

This tutorial gives a brief overview of gene trap technology, reviews gene trap vector types and function, and discusses experimental opportunities available to gene trap cell line users. For information on how to locate a gene trap cell line on the IGTC web site, please see our search tutorial.

- [Introduction to Gene Trapping](#)
- [Vector Types and Function](#)
- [Experimental Opportunities](#)

Introduction to Gene Trapping

Gene trapping is a method of randomly generating embryonic stem cells with well-characterized insertional mutations. The mutation is generated by inserting a gene trap vector construct into an intronic or coding region of genomic DNA. The gene trap vector constructs contain selectable reporter tags used to identify cell lines where the vector has successfully interrupted a gene. These reporter tags can also be useful for further experimentation in cells and mice. Gene trap sequences are derived from cDNA or genomic DNA from the trapped locus using primer sequences from vector ends, and the sequences are used to identify and annotate the trapped gene. Gene trap cell lines reliably contribute to the germ line, producing very useful mutant mouse strains for the functional characterization of genes. Although the insertion of the vector construct in a gene region typically results in complete inactivation of the "trapped" gene (a null allele), this is not guaranteed. In some cases vector insertion can fail to inactivate a gene, lead to hypomorphic gene function, or result in a dominant negative phenotype. Generally, vector insertion close to the 5' end of a gene, but downstream of the untranslated region before the first exon, is more likely to create a null allele than insertion near the 3' end.

The International Gene Trap Consortium website represents all publicly available gene trap cell lines, which are distributed on a non-collaborative basis for nominal handling fees. By using gene trap cell lines found on the IGTC site, researchers can save the time and expense of targeting a gene for knockout. Researchers can find trapped genes of interest on the IGTC website (see tutorial on finding gene trap cell lines of interest), and have the cell lines sent to their lab for the generation of mutant mice through blastocyst injection.

Vector Types and Function

Gene trap vectors are designed to insert into genomic sequence and interrupt transcription of the trapped gene. There are a variety of different gene trap vector types, and each will produce cell lines with different characteristics and research opportunities. Researchers are advised to learn the characteristics of the different vectors used to create cell lines available for a particular gene or locus of interest. This information is available on the IGTC site on the cell line annotation page and on IGTC member websites. For a more detailed review of gene trap technology see:

Stanford, W.L., Cohn, J.B. & Cordes, S.P.
Gene-trap mutagenesis: Past, present and beyond.
Nature Reviews Genetics 2, 756-768 (2001).

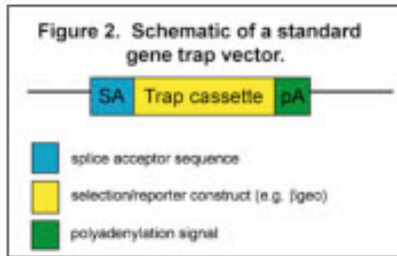
Figure 1 shows how genes are normally transcribed and spliced into mRNA products. Gene trapping takes advantage of the splicing apparatus by using a vector construct containing a splice acceptor signal, causing the vector sequence to be spliced into the mRNA. Gene trap vectors contain a polyadenylation signal at the 3' end that causes the mRNA to be truncated and non-functional.

Figure 1. An endogenous promoter drives transcription of a gene, which is followed by normal splicing.

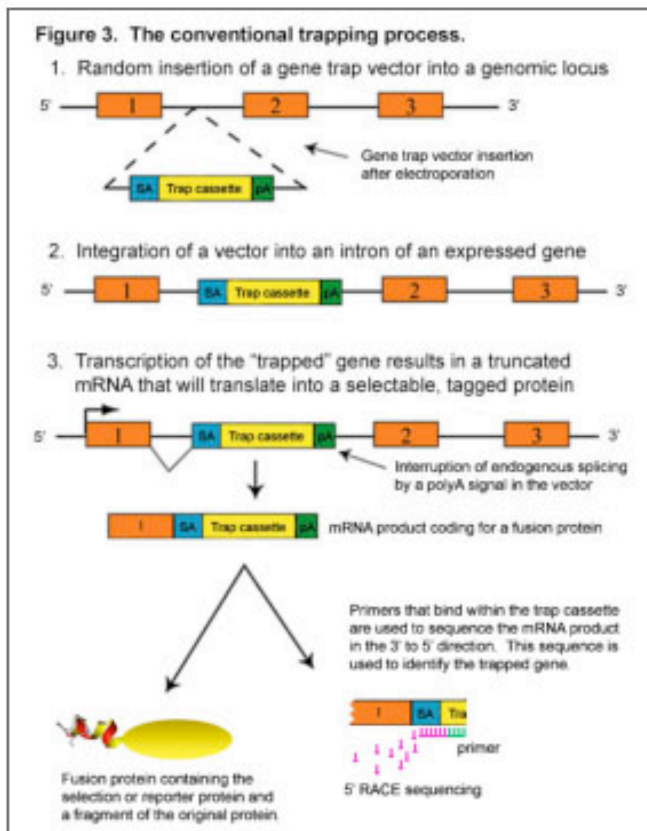
The diagram shows a DNA strand with a promoter (indicated by an arrow) and three exons labeled 1, 2, and 3. An arrow points down to a pre-mRNA molecule with introns. A second arrow points down to a mature mRNA product consisting of exons 1, 2, and 3 joined together. A final arrow points down to a 3D protein structure labeled 'protein product'.

The basic traits of a gene trap vector are shown in Figure 2 below. The splice acceptor interrupts normal splicing and causes the downstream vector sequence to be transcribed. The gene trap cassette contains a combination of selection and reporter constructs and is followed by a polyadenylation signal, which causes a stop in translation. PolyA vectors work in a different manner, using a promoter and a splice donor to trap the 3' ends of genes, shown in more detail later.

31



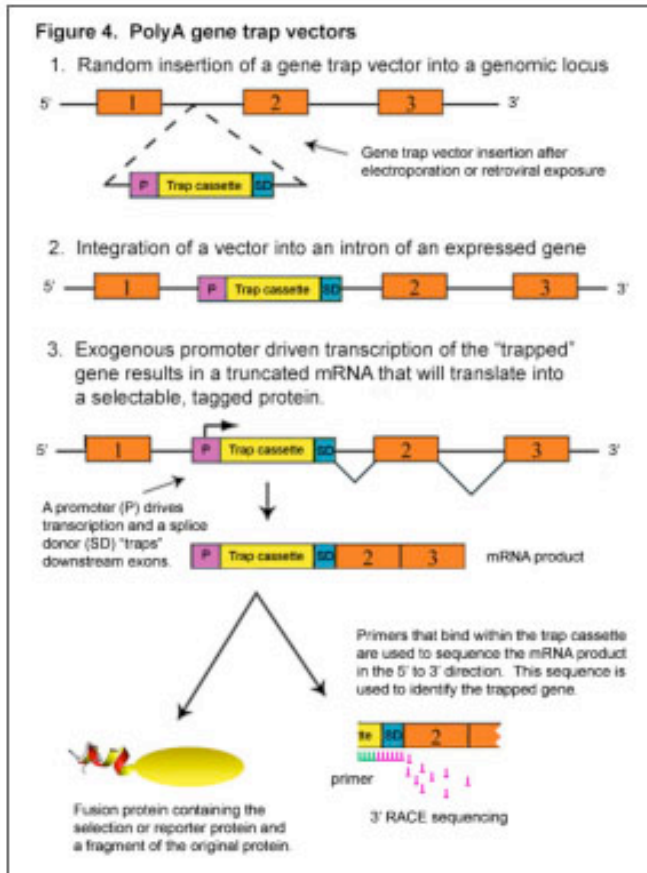
There are two main strategies employed by gene trap vectors: conventional vectors use the endogenous promoter, while polyA vectors contain a strong promoter in the vector sequence. Conventional gene trap vectors use a splice acceptor to take advantage of endogenous transcription and truncate the mRNA, leaving the gene 5' of the insertion site intact, followed by the vector sequence containing the selection/reporter construct. A polyA signal is placed at the 3' end of standard vectors, causing translation to end and producing a truncated fusion protein. This is shown in figure 3 below.



PolyA gene trap vectors employ a different strategy, shown in figure 4. These vectors contain a promoter signal and a transcriptional start site, allowing genes to be trapped that are not normally expressed, or are expressed at very low levels under experimental conditions. A splice donor sequence is present at the end of the gene trap cassette, causing the mRNA product of the vector construct to be fused with any downstream exons. Since these vectors do not have their own polyA sequences to signal the end of translation, only cell lines in which the vector inserts upstream of a terminal exon will produce the selectable reporter tag.

It is important to note that the sequence used to identify a gene inactivated by a polyA gene trap vector is taken from exons 3' of the vector insertion site. This differs from typical gene trap sequence which is taken from the exons 5' of the vector insertion site.

Figure 3.4. Page 2 of the IGTC Overview Tutorial web page.



Experimental Opportunities

Gene trap vectors produce different insertional mutations in genes, resulting in different allele types and different options for further manipulation of the trapped locus. Conventional gene trap vectors will produce a null fusion protein that is regulated in the same manner as the trapped gene. PolyA vectors introduce a promoter in the locus to drive transcription, creating constitutively active transcription of the mutant protein. In addition, the insertion site of the gene trap vector will affect the protein as well, with some truncated proteins retaining partial functionality depending on the intragenic location of functional domains.

Newer gene trap vectors have been developed that incorporate site-specific recombination sites, allowing for further modification of the trapped locus. These sites may be used to create different alleles, including reverting the trapped gene to wild type, and creating conditional alleles. Figure 5 shows schematic examples of how these systems work. For more detailed information on the characteristics and options available for particular vectors, view the IGTC publication list and visit IGTC member websites. For more information about post-insertional modification and site-specific recombination, see:

Brandt, C.S. & Dymecki, S.M.
Talking about a revolution: The impact of site-specific recombinases on genetic analyses in mice.
Developmental Cell 6, 7-28 (2004).

Figure 3.4. Page 3 of the IGTC Overview Tutorial web page.

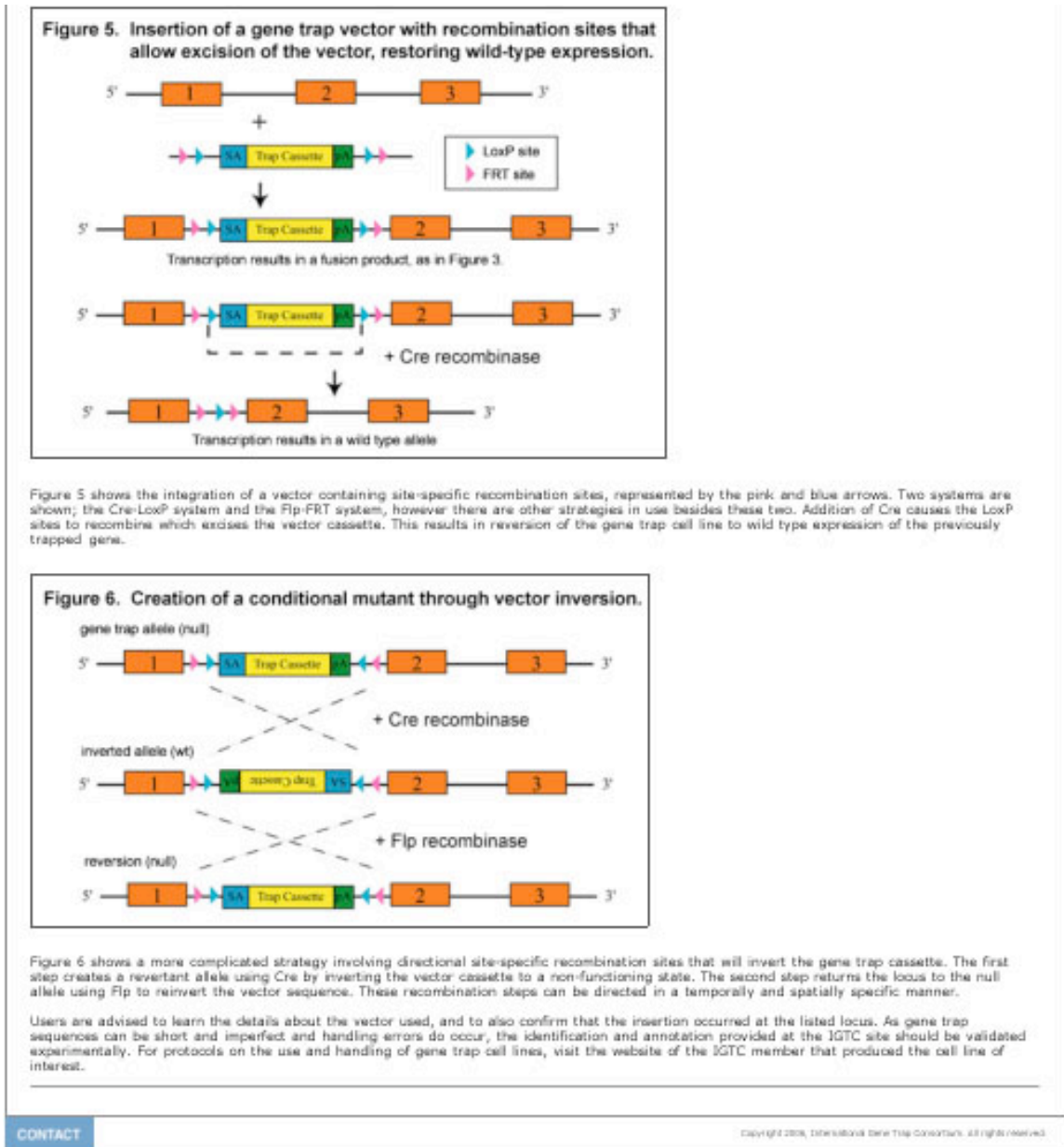




Figure 3.4. Page 4 of the IGTC Overview Tutorial web page.

Figure 3.5. The following 14 pages contain the contents of the IGTC Search Tutorial, which displays information about gene trapping. Below is the first page of the tutorial.


International Gene Trap Consortium


INFORMATION
DATA ACCESS
TUTORIALS
REQUEST ES CELL LINES

Tutorials: Locating A Cell Line


Locating a gene trap cell line in a gene/locus of interest

This tutorial demonstrates how to locate a gene trap cell line in a gene/locus of interest. For information on the gene trapping process, please see our overview tutorial.

The best way to determine if the IGTC has trapped your gene or locus of interest is to use the BLAST search function to align your sequence to our database of trapped genes and cell line sequences. If you do not know the sequence of your gene, you may perform searches based on keywords or expression profile, or browse the contents of the IGTC database.

- [BLAST Search](#)
- [Browse the IGTC](#)
- [Search the IGTC](#)
- [Browse Biological Pathways or Gene Ontology categories](#)
- [Search expression data](#)
- [Browse Ensembl or UCSC Genome Browsers](#)

To begin a search, select [Data Access](#) in the menu bar at the top of the page. Scroll down to select the type of search you wish to perform.


International Gene Trap Consortium

INFORMATION
DATA ACCESS
TUTORIALS
REQUEST ES CELL LINES

[About IGTC](#)

[Gene trapping mutations and generating cell lines](#)

[Keyword/ID Search](#)

[Blast Search](#)

[Expression Search](#)

[Browse](#)

[Biological Pathways](#)

approach that is used to introduce insertional mutations into embryonic stem (ES) cells. In addition to random insertions, newer gene trap vectors offer a variety of post-insertional modification strategies for the generation of other experimental alleles.

BLAST Search

Sequence searching with BLAST is the best way to find matches in the IGTC database. The BLAST results will find all matches with significant sequence similarity, regardless of annotation.

To start your BLAST search, select the [Blast Search](#) option from the [Data Access](#) menu located on the menu bar. It will take you to the following BLAST form:

Blast

Blast

A local BLASTN search will be run on the IGTC databases using NCBI's BLAST server software

Choose the database to search:

cell line tags

Enter sequence below in FASTA format

```
>NM_146108 (Nabch)
GTCTAACTACCGAGCCGGTGGAAACCACTACACCTCTGCTTCTTCTGCTCTTGGGTTTTTMM
CAGCCATATGCGTGGCCGCTCCTGTCSAGGTCAGCTCCTTCASAAGAGCCAGCCGTATT
TGAGAATGTCCATGCACACAGAGCTGCAGAACTGCTGCTGGAGAGAGAGGCTGTGGAG
GCTCAACAGACCCAASTTCTCAACGCACTGASTCTGAATATGATCCGGCAGATCTATCC
ACATGGGAACAGACCCCTGACACATTCTGTATCATATAAAGGGAGCCGGAGAAAGGCC
```

Clear sequence Quick Search

A Quick search is run with the default settings listed below.

The query sequence is filtered for low complexity regions by default.
Filter Low complexity

Expect 10

Graphical Overview Alignment view Pairwise

Descriptions 50 Alignments 30

Clear sequence Search

Sample search results are available.
This interface was created by Sergei Shavirin at the National Center for Biotechnology Information.

Enter your the sequence of your gene of interest in *Fasta format* and dick on the **Quick Search** button. Multiple sequences can be searched simultaneously, with results appearing in the same order the sequences were entered in the search field.

You can change the settings of your BLAST search if you do not wish to use the default settings. You can turn on or off low complexity filtering, which prevents matches to regions with highly repetitive sequence. You can increase or decrease the maximum allowable expect value, a term inversely correlated to the significance of a match, with the smallest numbers indicating the best matches. To alter the presentation of the results, you can turn on or off the graphical overview, increase or decrease the number of descriptors and alignments shown, or change the alignment view. See the NCBI BLAST Handbook or FAQ page for further details. Once you have made the adjustments to the settings, click on the **Search** button.

The default BLAST search for the NM_146108 sequence yields the following result: (alignments not pictured)

Figure 3.5. Page 2 of the IGTC Search Tutorial web page.

NCBI **BLAST Search Results** BLAST Entrez ?

BLASTN 2.2.5 [Nov-16-2002]

Reference:
 Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Database: igtc_cellline
 32,701 sequences: 8,518,844 total letters

Query= NM_146108 (Hiboh)
 [1720 letters]

Distribution of 15 Blast Hits on the Query Sequence

Mouse-over to show define end scores. Click to show alignments

Color Key for Alignment Scores

<0	40-50	50-80	100-200	>200
----	-------	-------	---------	------

Sequences producing significant alignments:	Score (bits)	E Value
RR5545 (BG)	1006	0.0
RR8007 (BG)	33	1.2
A08A02 (GGTC)	31	4.6
A057A01 (GGTC)	31	4.6
XL244 (BG)	31	4.6
XE403 (BG)	31	4.6
XE402 (BG)	31	4.6
A007E01 (GGTC)	31	4.6
XG558 (BG)	31	4.6
FBCRC-GT-S5-2G1 (FBCRC)	31	4.6
A033E05 (GGTC)	31	4.6
A034A01 (GGTC)	31	4.6
M053A02 (GGTC)	31	4.6
RRF190 (BG)	31	4.6
RR5111 (BG)	31	4.6

Done

The cell line RR5545 is the best IGTC match to NM_145108 (the gene of interest for this example). Record the cell line number. Search the IGTC for the cell line to get additional information from the cell line annotation page.

Browse the IGTC

You can browse the IGTC database for your gene(s) of interest. You can also browse through the available IGTC cell lines. Select *Browse* from the Data Access menu to start browsing the IGTC database.

Browsing by Gene

To start browsing by gene, click the circle next to Gene. The form will then give you options to sort the display of the available trapped genes:

Figure 3.5. Page 3 of the IGTC Search Tutorial web page.

Browse

Browse Form

Category: Gene Cell Line

Sort By: Gene Name

Status: --- N/A ---

Source: --- N/A ---

Entries Per Page: 1,000

Browse

You can sort the genes by

- **Gene Description** - Alphabetical order by gene description according to MGI. If a gene description is not available from MGI, the gene description is retrieved from either Entrez Gene or Ensembl.
- **Gene Symbol** - Alphabetical order by MGI gene symbol.
- **Chromosome** - Ordered by chromosome number.

You can also choose how many entries you want to show per page: 1,000, 5,000, or all entries. Once you have selected your parameters, click on the **Browse** button to start browsing.

Browse by Cell Line

To start browsing by cell line, click the circle next to Cell Line. The form will then give you options to sort the display of the available IGTC cell lines:

Browse

Browse Form

Category: Gene Cell Line

Sort By: Cell Line

Status: All

Source: All

Entries Per Page: 1,000

Browse

You can sort the cell lines by

- **Cell Line** - Alphabetical order by cell line name.
- **Source** - Alphabetical order by gene trap resource.
- **Chromosome** - Ordered by chromosome number.
- **Gene Description** - Alphabetical order by gene description according to MGI. If a gene description is not available from MGI, the gene description is retrieved from either Entrez Gene or Ensembl.
- **Gene Symbol** - Alphabetical order by MGI gene symbol.
- **Status** - Alphabetical by identification status (see below).

You can also limit the browsing by the status of the cell line and the source of the cell line.

To limit the cell lines by identification status, select one of the following status terms:

- **Localized** - Shows all cell lines where a single genomic locus has been identified by direct genome localization of the sequence tag derived from the cell line, genome localization of a full-length mRNA transcript related to the cell line, or both in agreement.
- **Conflict** - Shows all cell lines where the genomic loci identified by sequence tag localization and transcript localization do not overlap.
- **Unlocalized** - Shows all cell lines where there is no genomic locus found for either the cell line sequence or any transcript associated with the cell line sequence. Although there is no localization, there may be mRNA transcripts identified for an Unlocalized cell line.

To limit the browsing by gene trap resource, select one of the following genetrapp resources from the menu:

- **BG** - BayGenomics
- **CHRD** - Center for Modelling Human Disease
- **ESDB** - Embryonic Stem Cell Database
- **FHCRC** - Fred Hutchinson Cancer Research Center
- **GGTC** - German Gene Trap Consortium

Figure 3.5. Page 4 of the IGTC Search Tutorial web page.

- **SICTR** - Sanger Institute Gene Trap Resource
- **TIGEM-IRDM** - Telethon Institute of Genetics and Medicine - Istituto de Ricerche di Biologia Molecolare

You can also choose how many entries you want to show per page: 1,000, 5,000, or all entries. Once you have selected your parameters, click on the **Browse** button to start browsing.

Search the IGTC

You can search the IGTC database for a particular gene or cell line. Select **Keyword/ID Search** from the **Data Access** menu to start searching the IGTC database.

■ Search by Gene

To start searching by gene, click on the circle next to **Gene**. This will prompt the form to give you options to define the parameters for your search by gene:

You can search for a gene by

- **Keyword** - Any word, phrase, identifier, or number in any field of the database.
- **Accession** - The NCBI accession number: (e.g. NM_008255 or AK005015)
- **Gene Description** - A keyword search of the name or description of a gene. (e.g. 'ADP-ribosylarginine hydrolase' would be found with a full-name search or 'hydrolase', 'ADP', 'rbo', etc.)
- **Gene Symbol** - The MGI gene symbol (e.g. Adam23 or Cdad)
- **MGI ID** - The MGI gene identifier. (e.g. MGI:1345163)
- **Entrez ID** - The NCBI Entrez gene identifier. (e.g. 23792)
- **Ensembl ID** - The Ensembl gene identifier. (e.g. ENSMUSG00000025964)
- **Microarray** - A keyword search can be performed against the names of the probe sets in the IGTC database. However, if you do not know the name of the probeset, it is better to use the "search expression data" tool.
- **Phenotype** - A keyword search of phenotype information for a gene. This information is retrieved from the Mouse Genome Informatics web site.
- **Gene Ontology** - A keyword search of Gene Ontology terms associated with the gene. Further Gene Ontology information can be retrieved using the "BROWSE Biological Pathways or Gene Ontology categories" tool.

Once you select how you would like to search for your gene, enter the search term in the box labeled "Search Term".

You can limit your search by chromosome, by selecting a chromosome number in the Chromosome field.

You can also sort your search by

- **Gene Description** - Alphabetical order by gene description according to MGI. If a gene description is not available from MGI, the gene description is retrieved from either Entrez Gene or Ensembl.
- **Gene Symbol** - Alphabetical order by MGI gene symbol.
- **Chromosome** - Ordered by chromosome number.

Choose how many entries you want to show per page: 1,000, 5,000, or all entries. Once you have selected your parameters, click on the **Search** button to start your gene search.

■ Search by Cell Line

To start searching by cell line, click on the circle next to **Cell Line**. The form will give you options to define the parameters for your search by cell line:

Figure 3.5. Page 5 of the IGTC Search Tutorial web page.

You can search for a cell line by

- **Cell Line** - The cell line name (e.g. C91D-GT_BSAB-3 or K5T021)
- **Gene Description** - A keyword search of the name or description of a gene. [e.g. 'ADP-ribose/arginine hydrolase' would be found with a full-name search or 'hydrolase', 'ADP', 'ribo', etc.]
- **Gene Symbol** - The MGI gene symbol (e.g. Adam23 or Cxcl8)

Once you select how you would like to search for your cell line, enter the search term in the box labeled "Search Term".

You can limit your search by chromosome, by selecting a chromosome number in the Chromosome field.

You can sort your search by

- **Cell Line** - Alphabetical order by cell line name.
- **Source** - Alphabetical order by gene trap resource.
- **Chromosome** - Ordered by chromosome number.
- **Gene Description** - Alphabetical order by gene description according to MGI. If a gene description is not available from MGI, the gene description is retrieved from either Ensembl Gene or Ensembl.
- **Gene Symbol** - Alphabetical order by MGI gene symbol.
- **Status** - Alphabetical by identification status (see below).

You can also limit the search by the status of the cell line and the source of the cell line.

To limit the cell lines by identification status, select the one of the following status terms:

- **Localized** - Shows all cell lines where a single genomic locus has been identified by direct genome localization of the sequence tag derived from the cell line, genome localization of a full-length mRNA transcript related to the cell line, or both in agreement.
- **Conflict** - Shows all cell lines where the genomic loci identified by sequence tag localization and transcript localization do not overlap
- **Unlocalized** - Shows all cell lines where there is no genomic locus found for either the cell line sequence or any transcript associated with the cell line sequence. Although there is no localization, there may be mRNA transcripts identified for an Unlocalized cell line.

To limit the browsing by gene trap resource, select one of the following genetrapp resources from the menu:

- **BG** - BayGenomics
- **CHMD** - Center for Modeling Human Disease
- **ESDB** - Embryonic Stem Cell Database
- **FHCRC** - Fred Hutchinson Cancer Research Center
- **GGTC** - German Gene Trap Consortium
- **SGTR** - Sanger Institute Gene Trap Resource
- **TSGEM-IRBM** - Telethon Institute of Genetics and Medicine - Istituto de Ricerche di Biologia Molecolare

Choose how many entries you want to show per page: 1,000, 5,000, or all entries. Once you have selected your parameters, click on the **Search** button to start your cell line search.

Browse Biological Pathways or Gene Ontology categories

The IGTC, in collaboration with GenMAPP.org, has mapped gene trap data to sets of biological pathways and GO terms. You can browse the MAPPs (Map Annotator and Pathway Profiler) and Gene Ontology (Gene Ontology) Pathways by selecting *Biological Pathways* from the *Data Access* menu.

Once you select "Biological Pathways", you have a choice to view either MAPPs or Gene Ontology Pathways. Selecting MAPPs will allow you to view trapped genes in biological pathways. Selecting Go Pathways will allow you to view genes associated with GO terms.

When you look at the pathways, if a gene trap exists for a particular gene, the name of the cell line appears next to the gene. For example, in this portion of the Gene Ontology Term cell cycle MAPP, you can see the name of the gene trap (if available) appears next to the gene name. The colors indicate how many gene traps exist for that particular gene. See the legend for more detailed information.

Figure 3.5. Page 6 of the IGTC Search Tutorial web page.

Stag2	TEA112	Gene Database Mm-9st_20040824_Enr.qdb Expression Dataset Name: IGTC_20041009_trapset Color Set: Traps Gene Value: Cell Line ID Legend <input type="checkbox"/> No Traps <input type="checkbox"/> 1 Trap <input type="checkbox"/> 2 - 5 Traps <input type="checkbox"/> 6 - 10 Traps <input type="checkbox"/> 11-20 Traps <input type="checkbox"/> 21-50 Traps <input type="checkbox"/> > 50 Traps <input type="checkbox"/> No criteria met <input type="checkbox"/> Not found
Stag3		
Stk6	XG9148	
Tarpt	AC0203	
Tarf2	No Tra	
Tblp1	AO0723	
Tk1	CMHD-G	
Tik1	AP0206	
Tsc1	AO0182	
Tip53bp2	No Tra	
Ube2c	CMHD-G	
Zw10	No Tra	

Once you find the gene that you are interested in, click on the gene name to get more information. The gene annotation page will display information from MGI, SwissProt, Ensembl, Affymatrix, UniGene, RefSeq, Entrez Gene, and Gene Ontology. It will also have expression profile information. The Expression Profile section lists all the gene traps associated with the gene.

When you find the gene trap you would like to order, record the name. Search the IGTC for the cell line to get additional information from the cell line annotation page.

Search Expression Data

You can use the IGTC website to search for trapped genes that exhibit a desired expression profile in specific mouse tissues. The site supports two kinds of searches. The first is a search for genes that are upregulated/downregulated in a particular tissue. The second is a search for genes that have a specific expression level in a single tissue. Expression data is provided by the GNF SymAtlas project, and the search is designed to provide relative comparisons of expression levels, rather than analysis of statistical significance.

To begin a search of expression data, select Expression Search from the Data Access menu to start searching the IGTC database.

When the Search Form loads, in the Category field, select either 'Gene' or 'Cell Line'. A 'Gene' search will return genes associated with any expression data that meets your search criteria. Selecting 'Cell Line' will return cell lines that have been mapped to genes associated with expression data that meets your search criteria.

Next, choose the type of search you wish to perform.

Searching for upregulated/downregulated genes

Searching for genes that are upregulated/downregulated in a tissue allows the user to select the mouse tissue of interest, and the desired expression level for the trapped genes. The expression is calculated by comparing the expression level of each gene in the selected tissue with the median expression for the gene across all tissues. All genes that match the entered criteria are returned. By using this search, users can find trapped genes that are expressed in a tissue-specific manner.

To perform a search for upregulated/downregulated genes, select "Upregulated/downregulated" from the expression search page.

Searching by expression level

You can also search for all genes that match a desired level in a single tissue. This search compares gene expression levels in the tissue of interest to the median expression for all genes in the selected tissue. This search is useful for finding genes expressed or not expressed in a selected tissue. Again, the analysis is based on relative expression and is not meant to be statistically significant.

To perform a search by expression level, select "Expressed at a set level" from the search type menu.

Next, select a tissue to search. Keep in mind that, although the microarray data is extensive, not all genes have been tested in all tissues.

Then, choose an expression profile for which to search. In order to normalize across different microarray data sets, all expression levels are measured in comparison to the median expression level for a tissue or gene. Choosing "Above" will return any genes expressed at a level above the number you input times the median level for the tissue or gene. Likewise, choosing "Below" will return genes expressed at a level lower than your input number times the median. The input number can be any positive real number.

Finally, choose how you would like the results to be sorted, and select the number of results per page to be returned.

As an example, a search has been performed for genes that are upregulated or downregulated at 20 times the median level in dorsal root ganglia tissue, with the results sorted by gene description.

Figure 3.5. Page 7 of the IGTC Search Tutorial web page.

Search Expression Data

Search Form

Category: Gene Cell Line

Search Type: Upregulated/downregulated

Tissue: dorsal root ganglia

Expression Level: above 20 = Median

Sort By: Gene Name

Entries Per Page: 1,000

This search returned 13 genes in the IGTC database. The genes are organized by gene name or description, symbol and chromosome.

Viewing trapped genes that match the expression search criteria

Click on a gene name in the results, in the case of our example, "fibroblast growth factor 1". This will connect you to the IGTC Gene Annotation page for the appropriate gene.

Browse Genes With Expression Data

dorsal root ganglia: above 20 x median
Showing 1 - 13 out of 13 gene records

Description	Symbol	Chromosome
amyloid beta (A4) precursor-like protein 1	Aplp1	7
collapsin response mediator protein 1	Crmp1	5
DNA segment, Chr 5, Brigham & Women's Genetics 0060 expressed	D5Bwg0060e	5
ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)	Elavl2	4
fasciculation and elongation protein zeta 1 (zyg1)	Fez1	9
<u>fibroblast growth factor 1</u>	Fgf1	18
mitogen-activated protein kinase 8 interacting protein 3	Mapk8ip3	17
reticulin 1	Rtn1	12
Rho GDP dissociation inhibitor (GDI) gamma	Arhgdg	17
RIKEN cDNA E130013N09 gene	E130013N09rik	2
stathmin-like 2	Stmn2	3
synaptotagmin 11	Syt11	3
tubulin, beta 3	Tubb3	8

Toggling the arrows beside the Affymatrix Probe Sets and GNF Probesets (or selecting "Show All" under Additional Information) will display the Affymatrix or Novartis microarray chips containing the selected gene.

Figure 3.5. Page 8 of the IGTC Search Tutorial web page.

Gene Annotation

Gene

MGI Symbol: Fgf1
 Name: fibroblast growth factor 1
 Synonyms: Fam, Fgfa, Dftrx, Fgf-1, fibroblast growth factor 1 (acidic)
 Entrez: 14164 Chr.18(-): 39282129-39361648
 Ensembl: ENSMUSG00000036585 Chr.18(-): 39282130-39361649

Additional Information

Show All (▼) / Hide All (▲)

- ▶ Accessions
- ▶ Protein
- ▶ PubMed
- ▶ Homology
- ▶ Gene Ontology
- ▶ InterPro
- ▶ Protein Family
- ▶ MGI Phenotype
- ▼ Affymetrix Probesets
 - GeneChip Array: MG_U74Av2
 Probe Set ID: 100494_at
 - GeneChip Array: Mouse430A_2
 Probe Set ID: 1450869_at
 - GeneChip Array: Mouse430_2
 Probe Set ID: 1423136_at
 - GeneChip Array: Mu11KsubA
 Probe Set ID: us7610_s_at
- ▼ GNF Probesets
 - [gnf1m11999_at](#)
- ▶ Cell Line

Sequence Alignment Image

Alignment image is not currently available.

Clicking on the probe set name will connect you to the Affymetrix or Novartis website associated with that gene. As an example, the probe set "gnf1m11999_at" has been chosen, with the resulting Novartis page shown below. Please see the Novartis Frequently Asked Questions page for details regarding the use of the site.

Figure 3.5. Page 9 of the IGTC Search Tutorial web page.

View/Do Annot Table

Options Download

Go


Render Horiz. Bar Chart

Preset Check All

Filter None [Add](#)

[Home](#) [Search Expression](#) [FAQ](#) [Terms of Use](#) [About](#)

Search



Genomics Institute of the
Novartis Research
Foundation

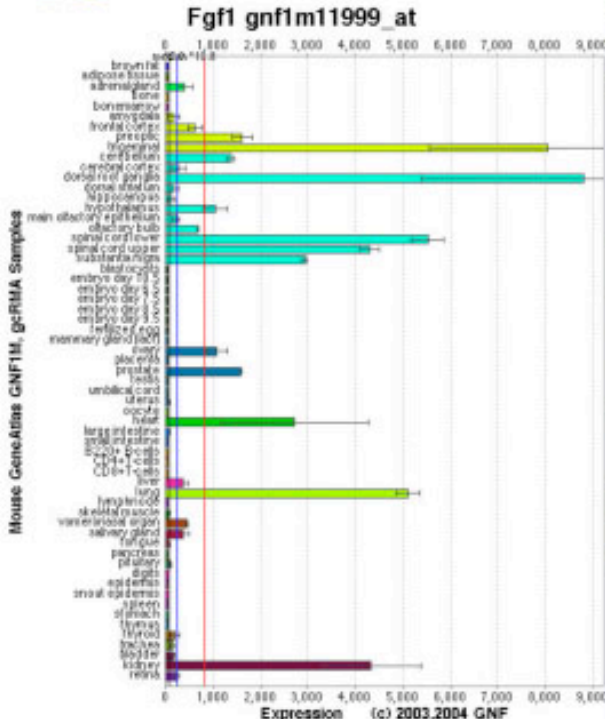
Fgf1 Dataset: Mouse GeneAtlas GNF1M, gcRNA Reporter: (all)

fgf1m11999_at in Mouse GeneAtlas GNF1M, gcRNA 0.9 Cut-off value

Search Results List

Mus musculus (1)

Fgf1



Fgf1 gnfm11999_at

Symbol	Fgf1
Description	fibroblast growth factor 1
Accession	U15102 (GenBank)
Aliases	Fgf1b , Fgf1 , Fgf1.1 , Fgf1a
Function	<ul style="list-style-type: none"> • cell differentiation (GO:0001954) • cell proliferation (GO:0002282) • development (GO:0007275) • fibroblast growth factor receptor binding (GO:0005104) • growth factor activity (GO:0005093) • heparin binding (GO:0005201) • induction of an organ (GO:0001759) • signal transduction (GO:0007165)
Protein Family	<ul style="list-style-type: none"> • Cytokine, IL-1 related (PRO00998) • Interleukin 10/heparin-binding growth factor (PRO02348)
Genome Location	<ul style="list-style-type: none"> • Chr 10:29,202,39,261 Mbp (c) NCBI M23 • Chr 10:29,052,39,142 Mbp (c) NCBI M24
Transcript	<ul style="list-style-type: none"> • ENSMUST00000040942 (Ensembl) FASTA (internal) • NM_031817.2 (RefSeq) SEQ COS FASTA (internal)
Protein	<ul style="list-style-type: none"> • ENSMUSP00000048710 (Ensembl) FASTA (internal) • NP_024327.1 (RefSeq) SEQ FASTA (internal) • P01146 (P/TREMBL) FASTA (internal)
Sequence Cluster	<ul style="list-style-type: none"> • Mm_2d1202 (UniProt)
Reporter	<ul style="list-style-type: none"> • gnfm11999_at (gnGNF1Musa) FASTA (internal) • 100401_at (M0_U744v2) FASTA (internal)

Browse Ensembl or UCSC Genome Browsers

You can also browse for a gene trap cell line using either the [Ensembl](#) or the [UCSC Genome Browsers](#).

▪ **Ensembl**

Go to <http://www.ensembl.org>, the Ensembl web site. Select the **mouse** genome. Once you are on the Mouse Genome Server page, shown below, either select a chromosome to browse through or enter a specific location on your chromosome of interest. A region on chromosome 13 has been entered as an example. Please click on the image to see a larger, clearer image.

Figure 3.5. Page 10 of the IGTC Search Tutorial web page.

The screenshot shows the Mouse Genome Server interface. At the top, there are logos for e! MGSC Mouse, the Wellcome Trust Sanger Institute, and EBI. Below the logos is the title "Mouse Genome Server".

The main section is titled "Ensembl Entry Points" and contains a search form with the following fields and options:

- Search for:** A dropdown menu set to "Anything" and a text input field.
- with:** A text input field.
- Display Chr:** A dropdown menu set to "13".
- From:** A text input field containing "91120468".
- To:** A text input field containing "91122532".

There are two "Lookup" buttons on the right side of the search form. Below the search form are two buttons: "Retrieve a sequence" with an "Export" button next to it, and "Search your sequence" with a "BLAST-SSAHA" button next to it. To the right of these buttons is the text "Advanced data retrieval tool" with a "Ensembl" button.

Below the search form are two columns of content:

- About MGSC:** A section with a mouse icon and text describing the Mouse Genome Sequencing Consortium as a joint project between The Wellcome Trust Sanger Institute, The Washington University Genome Sequencing Center, and others. It includes a link to "details".
- Mouse Genome Assembly NCBI m33:** A section with text stating that the site provides a full Ensembl gene build for the NCBI m33 mouse assembly (freeze May 27, 2004, strain C57BL/6J). It mentions extensive QC from both the Sanger Institute and NCBI to address assembly issues.
- Browse a Chromosome:** A section showing a set of 22 chromosome ideograms, numbered 1 through 11, 12 through 19, and X and Y.

At the bottom of the page, there is a paragraph of text: "To view IGTC gene traps on a region of the genome, scroll down to the Detailed view section, and select GeneTrap from the list of DAS sources as shown below. Please click on the image to see a larger, clearer image."

Figure 3.5. Page 11 of the IGTC Search Tutorial web page.

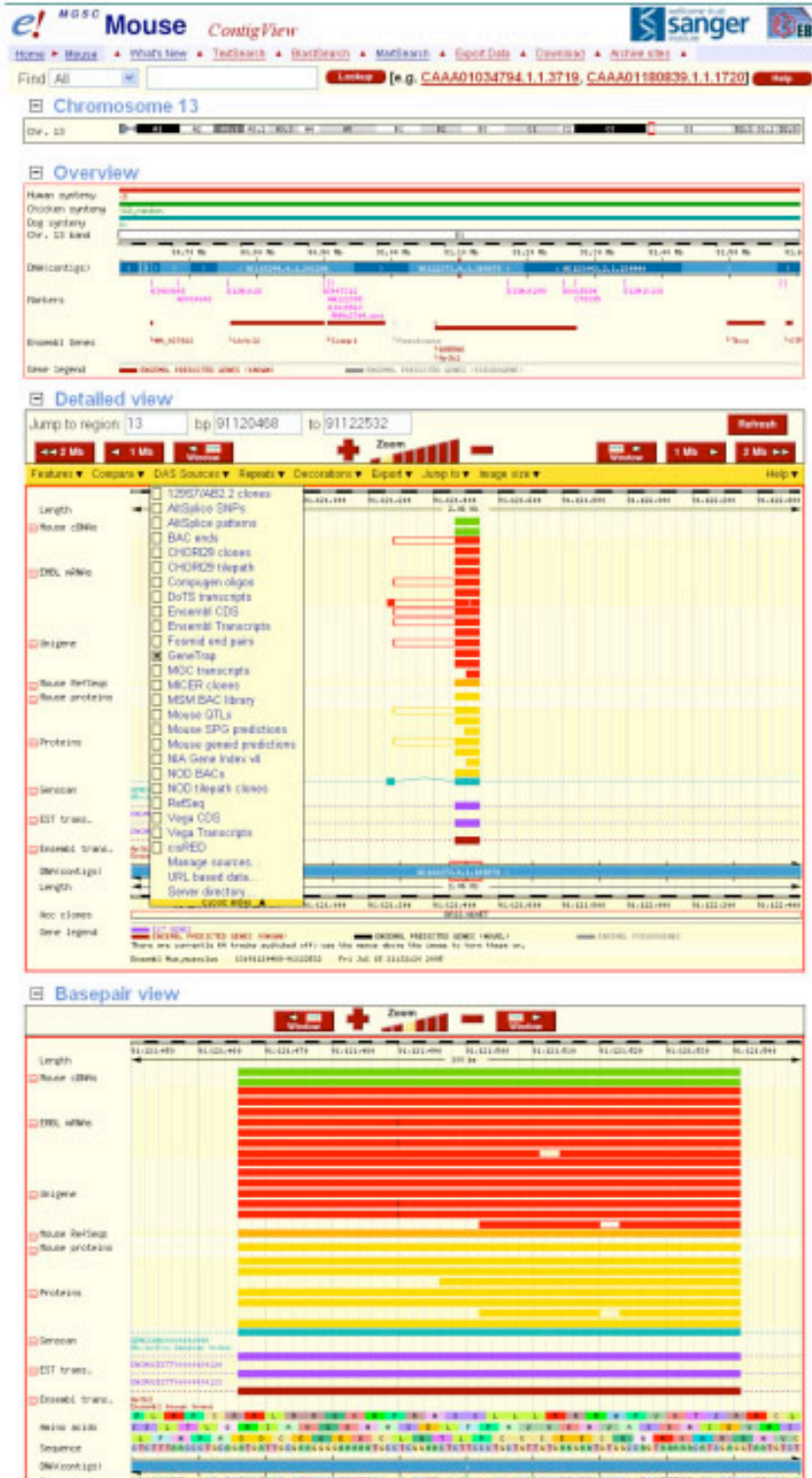
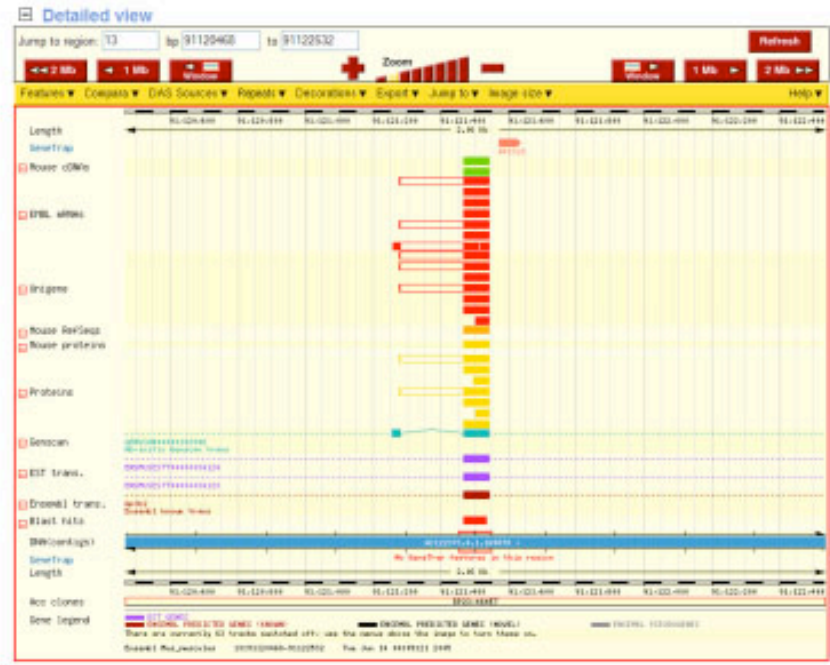


Figure 3.5. Page 12 of the IGTC Search Tutorial web page.

After selecting GeneTrap from DAS sources, click on the "close menu" option at the bottom of the menu. The page will automatically reload and the GeneTrap track will appear at the top of the **Detailed view** section.

Here is the same location as shown previously with the GeneTrap track visible:
Please click on the image to see a larger, cleaner image.



In this example, there is one gene trap call line available for this particular area of the genome. Click on that call line (RR1328 in this example) and select "DAS LINK:Genetrapp info" to get more information about the gene trap. It will take you to the annotation page of the IGTC resource that created the gene trap. You can also get more information on that gene trap by recording the call line name and searching the IGTC for the call line.

■ **UCSC Genome Browser**

Go to <http://genome.ucsc.edu>, the UCSC Genome Browser web site. Select **Custom Tracks** from the side bar on the left. Currently, only the subset of IGTC sequence tags that originate from BayGenomics tags are represented. Once you are on the Custom Annotation Tracks page, scroll down to the **Mouse Genome** section, and look for the BayGenomics listing. Each chromosome is listed separately, but clicking on any of them will activate the BayGenomics custom track for the whole genome. As chromosome Y is the smallest, it will be the quickest to load. Pictured below is the UCSC mouse genome browser, showing the BayGenomics knockout sequence tags on the Y chromosome.

Figure 3.5. Page 13 of the IGTC Search Tutorial web page.



Figure 3.5. Page 14 of the IGTC Search Tutorial web page[10].

References

1. Stanford WL, Cohn JB, Cordes SP: **Gene-trap mutagenesis: past, present and beyond.** *Nat Rev Genet* 2001, **2**(10):756-768.
2. Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L *et al*: **BayGenomics: a resource of insertional mutations in mouse embryonic stem cells.** *Nucleic Acids Res* 2003, **31**(1):278-281.
3. **Program in Genomic Applications, funded by the National Heart, Lung, and Blood Institute (NHLBI:U01 HL66600-01). This program is a major initiative to advance functional genomic research related to heart, lung, blood, and sleep health and disorders.** In.
4. Nord AS, Chang PJ, Conklin BR, Cox AV, Harper CA, Hicks GG, Huang CC, Johns SJ, Kawamoto M, Liu S *et al*: **The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse.** *Nucleic Acids Res* 2006, **34**(Database issue):D642-648.
5. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29**(1):11-16.
6. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, Anagnostopoulos A, Baldarelli RM, Baya M, Beal JS, Bello SM *et al*: **The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology.** *Nucleic Acids Res* 2005, **33**(Database issue):D471-475.
7. Cox T NA: **MAPTAG**
<http://www.sanger.ac.uk/PostGenomics/genetrap/maptag.shtml>. In.; 2003.
8. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser.** *Genome Res* 2004, **14**(5):951-955.
9. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
10. Harper CA, Huang CC, Stryke D, Kawamoto M, Ferrin TE, Babbitt PC: **Comparison of methods for genomic localization of gene trap sequences.** *BMC Genomics* 2006, **7**:236.

Introduction to the manuscript

“Comparison of methods for genomic localization of gene trap sequences”

One of the top priorities of the IGTC development team was to develop a sequence identification protocol that incorporated the best parts of the identification methods used by the different gene trap researchers joining the IGTC. After concluding that a two-path identification protocol would be used to identify sequence tags by alignment with gene transcripts and alignment with the mouse genome, a genomic localization method was needed. In order to determine which genomic localization program would be best suited for localizing the IGTC sequence tags to the mouse genome, I compared the performance of programs in use at the major genome browser web sites. The other members of the IGTC development team, Conrad Huang, Doug Stryke, Michiko Kawamoto, Thomas Ferrin, and Patricia Babbitt, also made contributions to this research. We produced some interesting conclusions that merited publication in *BMC Genomics*. The published manuscript below is presented in the following chapter.

Harper CA, Huang CC, Stryke D, Kawamoto M, Ferrin TE, Babbitt PC:
Comparison of methods for genomic localization of gene trap sequences. *BMC Genomics* 2006, **7**:236.

The nucleotide sequences used in this analysis are available from the BMC Genomics web site (<http://www.biomedcentral.com/bmcgenomics>) in association with this manuscript. The titles of these files, and their contents, are as follows:

Title: Sequence tags

File format: FASTA

Description: A file of sequence tags aligning to known genes that were used in “Comparison of methods for genomic localization of gene trap sequences”. This is a smaller set of sequences than is contained in the International Gene Trap Consortium database (<http://www.genetrap.org>).

Title: Genes

File format: FASTA

Description: A file of full-length genes aligning to the sequence tags that were used in “Comparison of methods for genomic localization of gene trap sequences”.

Comparison of methods for genomic localization of gene trap sequences

Courtney A. Harper¹, Conrad C. Huang², Doug Stryke², Michiko Kawamoto², Thomas E. Ferrin^{1,2}, Patricia C. Babbitt^{§,1,2}.

¹Department of Biopharmaceutical Sciences, University of California San Francisco, 1700 4th Street, San Francisco, CA 94143-2250, USA.

²Department of Pharmaceutical Chemistry, University of California San Francisco, 1700 4th Street, San Francisco, CA 94143-2250, USA.

[§]To whom correspondence should be addressed. Tel: +1 4154763784; Fax: +1 4155144260;

Email addresses:

CAH: courtney.harper@ucsf.edu

CCH: conrad@cgl.ucsf.edu

DS: stryke@cgl.ucsf.edu

MK: michiko@cgl.ucsf.edu

TEF: tef@cgl.ucsf.edu

PCB: babbitt@cgl.ucsf.edu

Abstract

Background

Gene knockouts in a model organism such as mouse provide a valuable resource for the study of basic biology and human disease. Determining which gene has been inactivated by an untargeted gene trapping event poses a challenging annotation problem because gene trap sequence tags, which represent sequence near the vector insertion site of a trapped gene, are typically short and often contain unresolved residues. To understand better the localization of these sequences on the mouse genome, we compared stand-alone versions of the alignment programs BLAT, SSAHA, and MegaBLAST. A set of 3,369 sequence tags was aligned to build 34 of the mouse genome using default parameters for each algorithm. Known genome coordinates for the cognate set of full-length genes (1,659 sequences) were used to evaluate localization results.

Results

In general, all three programs performed well in terms of localizing sequences to a general region of the genome, with only relatively subtle errors identified for a small proportion of the sequence tags. However, large differences in performance were noted with regard to correctly identifying exon boundaries. BLAT correctly identified the vast majority of exon boundaries, while SSAHA and MegaBLAST missed the majority of exon boundaries. SSAHA consistently reported the fewest false positives and is the fastest algorithm. MegaBLAST was comparable to BLAT in speed, but was the most susceptible to localizing sequence tags incorrectly to pseudogenes.

Conclusions

The differences in performance for sequence tags and full-length reference sequences were surprisingly small. Characteristic variations in localization results for each program were noted that affect the localization of sequence at exon boundaries, in particular.

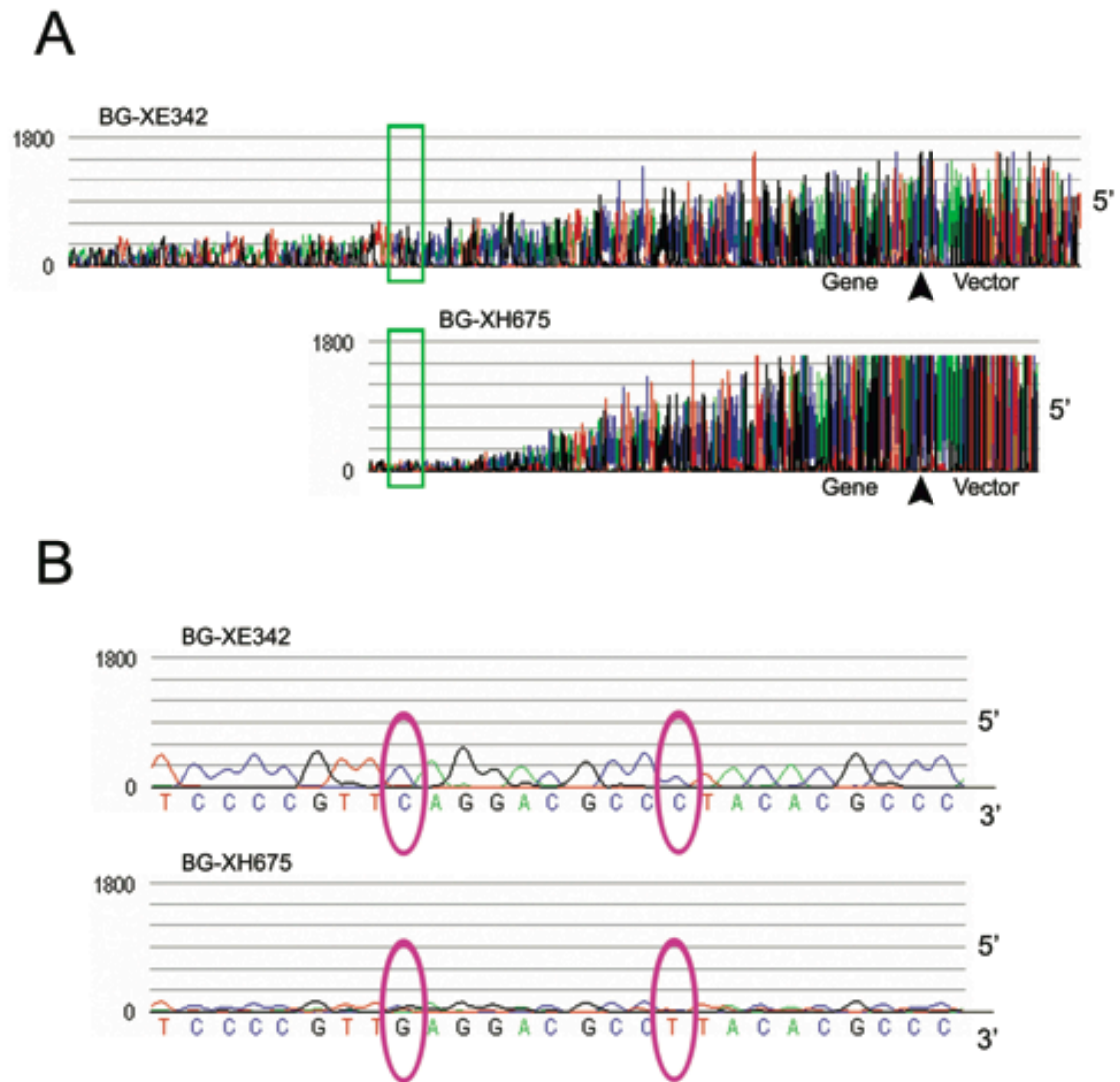
Background

High-throughput gene interruption projects have greatly increased the number of loss-of-function knockout genes available for study [1]. Correct identification of these genes provides a necessary foundation for their use for biomedical discovery, including minimizing the number of time-consuming phenotype experiments that need to be undertaken. Until recently, interrupted knockout genes have been identified primarily using the alignment program BLAST [2] to match gene trap sequence tags, which represent the region of an interrupted gene near the site of disruption, with gene transcripts. While transcript identification can generally provide high confidence gene annotation information for over 75% of such knockouts [3], transcript databases do not provide full coverage of the genome, limiting the number of genes that can be identified. Redundancy in transcript databases also makes it difficult to obtain a unique identification for sequence tags, which are relatively short.

Sequence quality can also be an issue with gene trap sequence tags, since the prevalent method of generating these tags often results in relatively low-quality sequence. BayGenomics [3] and other members of the International Gene Trap Consortium (IGTC)

[4, 5] typically use 5' RACE [6], a common method for amplifying sequence from gene insertion events. This method generates sequence from only one strand of DNA, and often generates only relatively short sequences, with sequencing errors accumulating especially towards the 3' end. To obtain sequence tags that are sufficiently long to uniquely identify most genes, BayGenomics, for example, uses a limit for the acceptable quality of a base call that is lower than the generally accepted threshold (a Phred [7] minimum score of 14.6 rather than the default score of 30) [3]. The consequence of using such a low threshold is that nucleotides are assigned incorrectly somewhat more often than with the default threshold value. This problem can interfere with annotation. [see Additional Figure 1 for an example.] Additionally, sequence tags generated by 5' RACE occasionally have non-templated nucleotides at their termini [8]. In one large-scale 5' RACE experiment, only 57% of clones generated sequences that were sufficiently long and unambiguous to be identified by alignment with a gene transcript [9].

As curation of the mouse genome has improved, direct localization has become the strategy of choice for associating sequence tags with specific genes. This has an advantage in minimizing imprecise and confusing annotations arising from redundancy in mRNA databases. Moreover, this approach reflects the biological reality of the insertion of a reporter gene into genomic sequence and provides a more context-based view of the gene by associating it with the many types of information available at the genome browser Web sites.



Additional Figure 4.1. An example of errors associated with low signal strength in a 5' RACE sequence. (A) Alignment of trace files for the sequence tags BG-XE342 and BG-XH675, both sequenced with 5' RACE, which localize to protein kinase C binding protein 1 (NCBI accession NM_027230). Black arrows indicate the point of vector insertion. The intensity of the signal diminishes towards the 3' end of each sequence. (B) Enlargement of the green-highlighted regions in A. The reverse complement of the trace sequence, which corresponds to the sequence of the inactivated gene, is listed below the expanded trace plots. The low intensity of the signal in this region of the BG-XH675 trace plot results in two nucleotide assignments, circled in pink, that differ from both genomic sequence from chromosome 2 and the associated mRNA sequence for this gene. In contrast, the corresponding nucleotide assignments in the relatively higher quality BG-XE342 trace plot, also circled in pink, agree with the genomic and mRNA sequences.

The choice of alignment program is a major consideration in localizing sequences on the genome. BLAST, which was developed for comparison of evolutionarily diverged sequences, is prohibitively slow in this application. Several newer algorithms have been developed to rapidly align nearly identical sequences. Implementations in common use are MegaBLAST [10], the Sequence Search and Alignment by Hashing Algorithm (SSAHA) [11], and the BLAST-like Alignment Tool (BLAT) [12]. Each is currently in use at one of the primary genome browser sites and, in addition, each is available for stand-alone use. MegaBLAST is used at the National Center for Biotechnology Information (NCBI) [13], SSAHA is used at Ensembl [14], and BLAT is used at the University of California Santa Cruz (UCSC) [15]. While all of these algorithms have been individually benchmarked for the genome browsers with which they are used, their performance with sequence tags has not been established, nor have the results from the stand-alone versions of these programs been compared with the gene annotations available at the genome browser sites. Establishing the effect of low quality and short sequence length on gene localization protocols is beneficial to research groups that work with gene tag and similar sequences, including other types of expressed sequence tags (ESTs) or genomic tags.

MegaBLAST is similar to BLAST in that it splits a query sequence into non-overlapping fragments and searches for exact matches to the genome to find the regions of highest identity. These perfect matches are then expanded to align the longest region of significant similarity. MegaBLAST uses a greedy algorithm that incorporates simplified gap and insertion/deletion penalties relative to BLAST and limits the number of alignments to be explored in extending the alignment beyond a perfect match seed.

These alterations are justified because of the high levels of similarity expected between query and database sequences and the expectation that the alignment will not contain many mismatches or gaps. For sequences with greater than 97% identity, MegaBLAST is an order of magnitude faster than BLAST without any loss of alignment accuracy [10].

SSAHA uses a different approach to take advantage of the high similarity expected between a query sequence and the genome. An index of all non-overlapping fragments of a set length (k) is created from the genome sequence and stored with the associated positions. The query sequence and its reverse complement are broken into all possible fragments of length k , including overlapping fragments, and compared with the genome index to identify exact matches. Matches are sorted to find contiguous matching segments that are reported if they exceed a threshold, set by default to $2k$. SSAHA is extremely fast, but due to the need to store the genome index and fragment locations, has relatively large memory requirements.

BLAT uses a multi-stage algorithm which searches for regions of similarity, aligns those regions, aggregates aligned regions in close proximity, and adjusts the boundaries of aligned regions to correspond with canonical splice sites. The initial search stage operates in a manner very similar to SSAHA. The genome database is broken into non-overlapping fragments of length k , then all k -length fragments of the query sequence and its reverse complement are associated with matching locations in the genome. The matches are sorted and grouped by proximity and those regions of the genome with a minimum of $2k$ contiguous matches are aligned with the query sequence. The alignment stage extends matching regions as far as possible, merges overlapping matches, links matches that fall in order on the genome into a single alignment, and fills in regions of

the alignment corresponding to gaps of identical length in the query and genome sequences. Positions of gaps in the alignment, which may correspond to introns, are matched to the consensus splice site GT/AG whenever possible.

The work reported here provides a comparison of the performance of the stand-alone versions of SSAHA, MegaBLAST, and BLAT for a set of mouse gene trap sequence tags. The sequence tags were generated through untargeted gene trap experiments, which detect instances where the insertion vector interrupts an intron of a gene expressed in embryonic stem cells [1]. As the genome coordinates of our sequence tags are not known, the localizations of their cognate genes were used as a proxy. These genes were identified by using the BLAST program to align the sequence tags with gene transcripts (see Methods for details).

The genome coordinates of many genes in the mouse genome are defined differently depending on which genome browser site provides the information. This is because each browser uses a different combination of localization programs, sequence analysis tools, and manual curation to arrive at their final annotations. Additionally, the localization program used in the annotation protocol may differ from the localization program provided to users of the genome browser. For example, Ensembl uses the exonerate program [16] to generate localization coordinates reported at their site. However, when a user seeks to localize a gene at the Ensembl site, the SSAHA algorithm is used to perform that task. This differs from NCBI and UCSC, where the localization algorithms used to generate annotations for the genome, MegaBLAST and BLAT respectively, are also used by the genome browser to localize sequences input by users. In order to provide a fair comparison between the algorithms, only sequence tags

matched with genes having exactly the same coordinates at Ensembl, NCBI, and UCSC were used in this study. To determine whether errors in the localization of sequence tags using the stand-alone versions of these programs was due to the nature of the sequence tags themselves or to differences in how the stand-alone programs perform relative to the protocol in which they are used to localize full-length genes at each browser site, we also localized the set of gene transcripts matched with sequence tags as a control. Our sequence set consisted of 3369 sequence tags associated with 1659 genes with uniformly assigned coordinates on the mouse genome.

Results and Discussion

Our results show differences in the localization performance with respect to recall and precision at each of three levels of granularity investigated, gene, exon, and nucleotide (Figure 1). The recall score indicates the percentage of true positives that were detected. Precision indicates the percentage of matches reported which correspond to true positives.

Localization to the correct gene

With respect to recall, our study shows that researchers who wish to link a sequence with information associated with the genome may confidently use any of the three localization programs considered in this study. SSAHA, MegaBLAST, and BLAT successfully localize each of the 1659 full-length genes in the test set to a genomic region

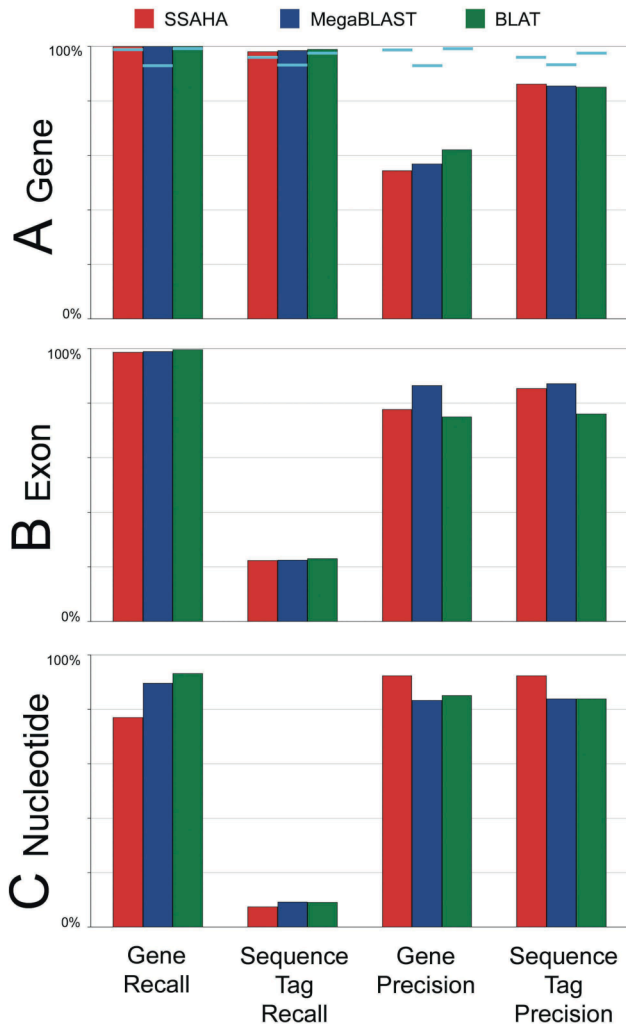


Figure 4.1. Recall and precision for each localization algorithm.

Results for SSAHA are shown in red, MegaBLAST in blue, and BLAT in green. The first column represents the recall obtained with full-length gene query sequences. The second column shows the recall obtained with sequence tag queries. The third and fourth columns display the precision of each algorithm when used to localize full-length genes and sequence tags, respectively. (A) Recall and precision at the level of the gene, as measured by overlap of at least one nucleotide between a set of localizations by an algorithm and the region of the genome containing the gene. Cyan lines indicate the recall and precision achieved when only the top hit is considered. (B) Exon recall and precision, as measured by an overlap of at least one nucleotide between the known localization of an exon and a match. Sequence tags are shorter than full-length genes and therefore typically contain sufficient sequence information to match only a few exons of any gene, leading to low recall at the exon and nucleotide levels. This does not indicate failure by the localization programs. (C) Nucleotide recall and precision, as measured by a match between a nucleotide in the known localization of a gene and a nucleotide from a query sequence localization.

that fully or partially matches the known coordinates of the corresponding gene (Figure 1A). Sequence tags fare nearly as well, with all programs reporting localization to the correct region of the genome for >98% of the 3369 sequence tags used in this study.

Repeat-masking of the genome accounts for the majority of the small number of failures in localizing sequence tags to the correct genes. Online localization is performed against masked genomic sequence by default as this ensures that results are returned quickly and that relatively few correct localizations are missed, despite the fact that as much as 50% of the genome consists of repeated elements [17]. In this study, less than 2% of sequence tags in the test set returned no localization results with one or more programs because they overlap fully or partially with regions removed by masking. Additionally, five sequence tags that localize to repeat regions have erroneous matches that exceed the minimum score required by each program, and so are localized incorrectly. In contrast, use of an unmasked version of the genome results in 100% recall for the test set of sequence tags, but increases the number of incorrect localizations by as much as ten-fold. Moreover, using an unmasked version of the genome increases computation time substantially (Table 1).

In contrast to the near-perfect recall exhibited by the localization programs, the precision of the programs suffers from a substantial incidence of false positives (Figure 1A). At the genic level, 46% of all reported full-length gene localizations and 16% of sequence tag localizations by SSAHA do not overlap with the known gene localization. For MegaBLAST, 43% of reported gene localizations and 15% of sequence tag localizations are false positives. BLAT shows similar performance, with 38% of reported gene localizations and 15% of sequence tag localizations falling outside the region of the

Table 4.1. Computation times in seconds for each algorithm.

		Computation Time in seconds		
	# of Sequences	MegaBLAST	SSAHA ^a	BLAT ^a
Full-length Genes	3320	1767 (40578) ^b	361 (29895)	1434 (204331)
Sequence Tags	7043	223 (1025)	38 (5806)	276 (854)

^a Reported computation times for SSAHA and BLAT do not include pre-indexing of the genome (see text).

^b Results using the repeat-masked genome are listed first, followed by results from the unmasked genome in parenthesis.

known gene. Generally, the false positives score significantly lower than the true positives.

False positives at the level of the gene may not be problematic, however, since the most common method of interpreting localization results is to accept the highest-scoring match as correct rather than analyzing all returned matches. Correct localizations generally exhibit long, high percent-identity matches, which contribute to higher scores compared with incorrect matches, which are generally short or contain mismatches. The strategy of taking the top hit is largely successful with both full-length gene queries and sequence tag queries (Figure 1A). The SSAHA localization with the highest score is almost always correct, as it overlaps with the known localization of a gene for 99% of full-length gene queries and 98% of sequence tag queries. The MegaBLAST localization with the highest score is correct for 93% of full-length gene queries, and 95% of sequence tag queries. The BLAT localization with the highest score is correct for 99% of full-length gene queries and 99% of sequence tag queries.

Erroneous matches are also less likely to group together on a chromosome than correct matches, which track with exon ordering. While all three programs report matches grouped by chromosome, only the BLAT algorithm incorporates matches in close proximity into a single multi-part alignment, which is given a score that combines the scores of the individual matches in the alignment. This ensures that the top-scoring match is a composite of all matches likely to be exons of the same gene. Another consequence of this grouping is that the scores of correct and incorrect matches are more widely separated than with SSAHA or MegaBLAST.

Pseudogenes

The presence of pseudogenes can confound rules for separating correct from incorrect matches at the genic level for both full-length genes and sequence tags. Pseudogenes are regions of the genome that are very similar in sequence to known genes, but are usually rendered non-functional by mutations or missing elements that prevent transcription or translation. About 80% of pseudogenes are processed pseudogenes, which resemble partial or full-length mRNA sequences that have been integrated into the genome [18]. These are caused by the retrotransposition of double-stranded DNA, read off of single-stranded RNA, into the genome. As processed pseudogenes lack introns, alignments can be constructed between pseudogenes and query sequences that are longer than individual exons. Such alignments may be sufficiently long that penalties accrued for mismatches are more than offset by this longer match length, allowing them to outscore correct matches to exons. In the case of our sequence tags, these alignments are invariably incorrect, since with our method of gene trapping, disruption of a gene is only detected when the vector is inserted into an intron [1]. Figure 2 gives an example that illustrates the difficulty in distinguishing localization to a processed pseudogene from localization to a true gene. More rarely, pseudogenes can be caused by duplications of chromosome segments. These unprocessed pseudogenes contain introns and are therefore less likely to result in high-scoring (but incorrect) matches based on alignment length alone. In addition, a recent duplication can result in a pseudogene with so few mutations that it may be difficult to distinguish it from the coding gene. Although it is possible for a gene trapping vector to insert into an unprocessed pseudogene containing

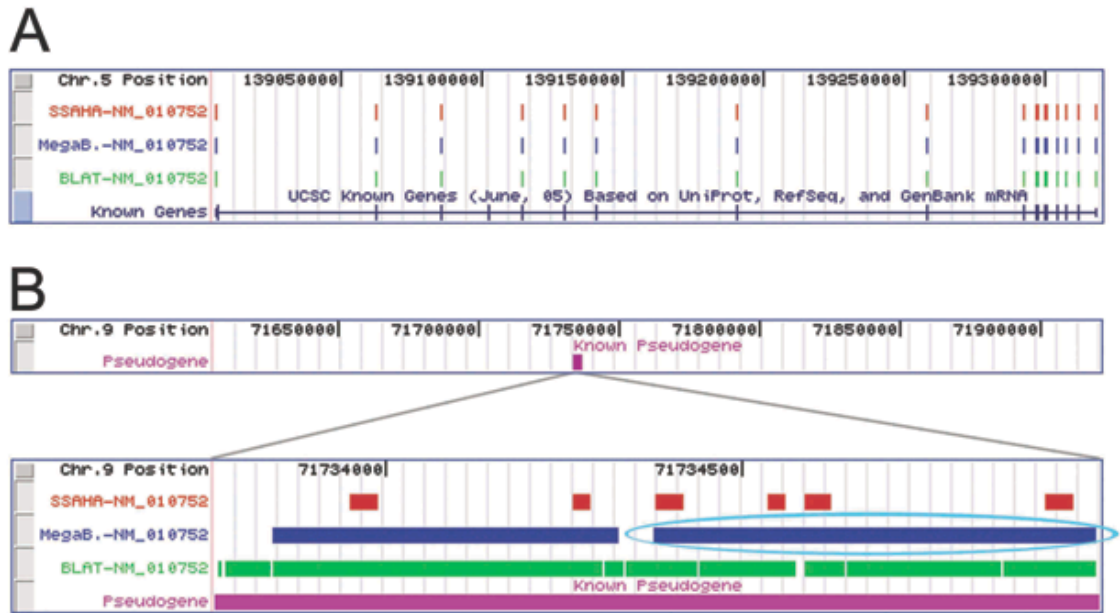


Figure 4.2. An example of localization to a pseudogene.

Localization results for the full-length gene encoding mitotic arrest deficient 1-like 1 (Mad11), GenBank accession NM_010752. All representations of alignments between query sequences and build 34 of the mouse genome were made using the UCSC Genome Browser Custom Tracks feature. Slight alterations have been made to the representations, including the removal of graphical elements to improve the clarity of the figure, but no changes were made to the alignments. (A) The coordinates of the known gene on the genome are listed at the top, and positions of exons are represented by colored blocks. A region of chromosome 5 is shown containing the known localization of NM_010752 (the Known Genes track at bottom) and the alignments of exons for NM_010752 to the genome by SSAHA, MegaBLAST, and BLAT. (B) A region of chromosome 9 containing a pseudogene related to NM_010752 is shown on the same scale as (A). Below this, the segment of chromosome 9 containing the pseudogene is enlarged. The highest-scoring MegaBLAST match, circled in cyan, localizes to this pseudogene rather than the real gene. The highest scoring matches returned by SSAHA and BLAT are located on chromosome 5 and overlap with the correct localization.

introns, none were detected in our data set, and thus all localizations to pseudogenes were considered false positives.

As shown in Figure 2, genic localization is compromised by the presence of pseudogenes to varying degrees. SSAHA identifies only exact matches, rather than very similar matches, lending the algorithm a distinct advantage in terms of distinguishing correct matches from pseudogene matches. BLAT alignments can contain mismatches accrued during the alignment extension stage, which increases the likelihood of a high-scoring match to a pseudogene. However, the BLAT score reflects all matches in a region of the genome so that short perfect or near-perfect exon matches in aggregate are likely to outscore longer imperfect matches to pseudogenes. MegaBLAST is the most susceptible to pseudogene matches, as it is relatively tolerant of mismatches and does not have a mechanism for favoring short perfect matches over long imperfect ones.

In this study, pseudogenes may have been the cause of over 100 top-scoring matches that are incorrect, despite high sequence identity between the query sequences and the genome. It is difficult to determine the exact number of incorrect localizations to pseudogenes as relatively few mouse pseudogenes have been annotated. As many as 4000 mouse pseudogenes are predicted to exist [19], and in the closely related human genome, a careful study of an early build of chromosome 22 revealed that 19% of sequences defined as coding likely belong to pseudogenes instead [20]. The distribution of pseudogene matches among the programs varies as might be expected from their algorithmic differences. SSAHA reports a top-scoring match to a region annotated as a probable pseudogene for 17 full-length genes and 60 sequence tags, while BLAT incorrectly localizes 7 genes and 45 sequence tags to probable pseudogenes.

MegaBLAST reports top-scoring matches to probable pseudogenes for 116 genes and 162 sequence tags.

Localization to the correct exon

With respect to recall, all three algorithms perform similarly well in localizing query sequences to the exons of their corresponding genes. For full-length gene queries, SSAHA, MegaBLAST, and BLAT all have exon recall of about 99% (Figure 1B). The sequence tags used in this study are generally substantially shorter than the full-length genes, averaging 255 nucleotides in length, versus 3611 nucleotides for genes, and it is rare that all exons of a gene will be matched in a sequence tag alignment. Thus, exon and nucleotide recall for sequence tag queries should be viewed in a comparative manner, rather than as a direct measure of the accuracy of each algorithm. SSAHA detects 22% of control exons, MegaBLAST detects 22% of control exons, and BLAT detects 23% of control exons.

Many of the exons that are not detected overlap with regions of the genome removed from the search space by repeat masking. Two examples of the effect of repeat masking on exon localization are illustrated in Figure 3, which depicts the genome alignment of the full-length gene encoding chromatin assembly factor 1, subunit A (Chaf1a), NCBI accession NM_013733, and the sequence tag BG-RRR265. Each program localizes NM_013733 to the left-most exon shown in Figure 3B by detecting perfect matches on either side of the repeat mask region. BLAT connects these matches because its default parameter settings allow alignments adjacent to a masked region to be extended into the masked sequence while SSAHA and MegaBLAST, whose default

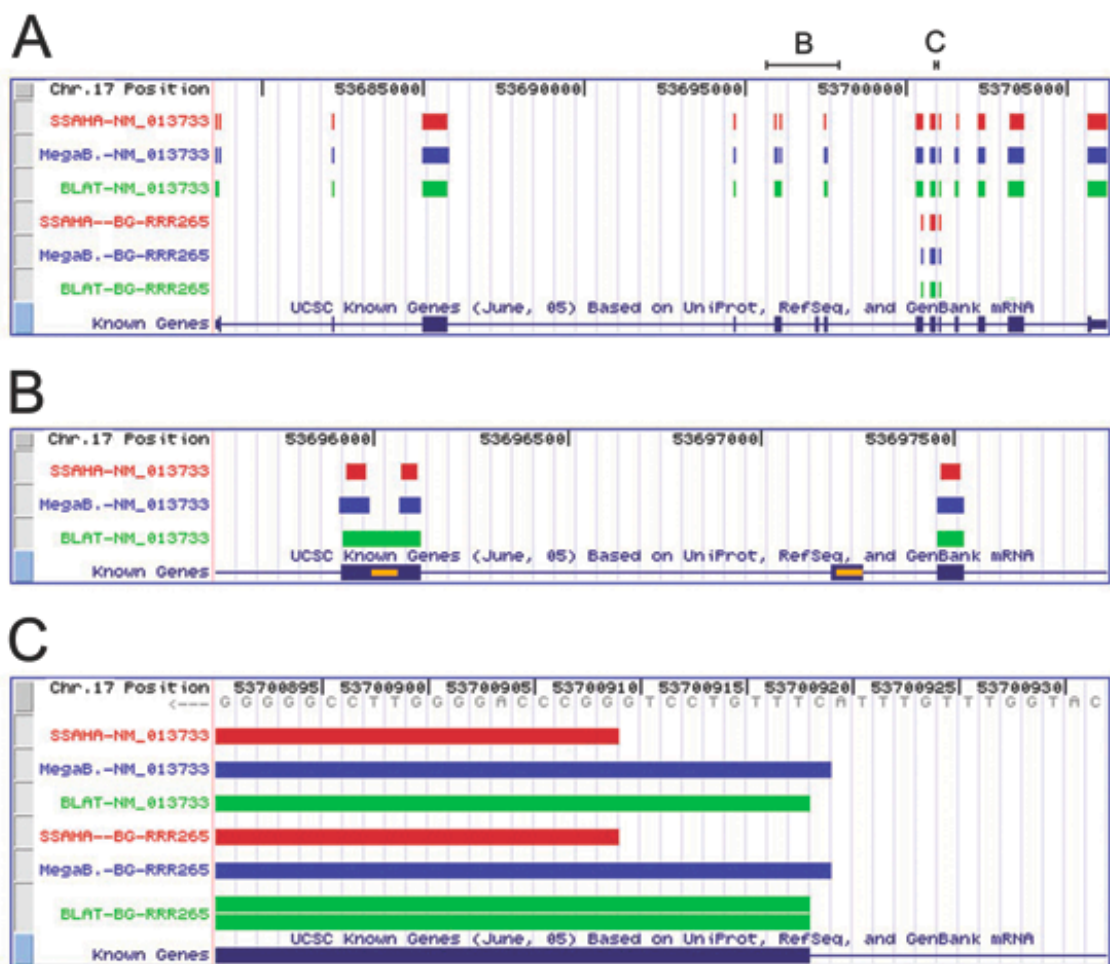


Figure 4.3. A representative genome alignment of a full-length gene and a sequence tag. The full-length gene encoding chromatin assembly factor 1, subunit A (Chaf1a), NCBI accession NM_013733, and the sequence tag BG-RRR265 align to a region of chromosome 17. (A) Overview showing the full region of the genome spanned by Chaf1a. Segments enlarged in the parts B-C are marked above the genome position. (B) Regions of genome that have been removed from the search space by repeat masking are shown in yellow, superimposed on the known gene track. The removal of these regions prevents correct localization of the full-length gene and sequence tag for these exons. (C) Magnification of the exon from region C illustrates differences between the alignment programs in aligning sequence to the edges of exons.

settings do not allow alignment to masked regions (see Methods), show the masked region as a gap. The middle exon in Figure 3B does not contain enough unmasked sequence for any algorithm to seed a match, and is thus entirely undetected.

Precision in exon localization is similar for full-length genes and sequence tags, despite the discrepancy in the number of exons matched by the two query types (Figure 1B). This indicates that although the short sequence tags do not contain regions matching all exons of their cognate genes, those with adequate length to be associated with a unique transcript generally contain sufficient information to be localized with high precision. For genes, 78% of SSAHA localization results overlap with known exons, compared with 85% for sequence tags. MegaBLAST has exon precision of 86% for genes, and 87% for sequence tags. BLAT has exon precision of 75% for genes, and 76% for sequence tags. Very rarely, the coding region of a gene contains an intron so short that MegaBLAST will align through it, including the intron in the alignment. This results in errors for four genes in the control set which contain introns of either 9 or 12 nucleotides in the upstream untranslated region. As a source of error, this had only a minimal effect on the overall precision for MegaBLAST and had no effect on the results for BLAT and SSAHA.

An interesting result that is not reflected by measures of recall and precision is that each program occasionally returns multiple correct localizations to the same exon. The full-length genes used in this analysis average 12.9 exons, but each program averages more than 13 correct localizations per gene. SSAHA returns 19.5 localizations per gene, with each localization corresponding to an exon or a false positive. On average, 15.2 of these aligned segments overlap with 12.8 exons. MegaBLAST returns 15.6

localizations per gene, with 13.5 of them correctly identifying 12.8 exons. BLAT returns 20.8 localizations per gene, with of them 15.6 correctly identifying 12.9 exons. Multiple localizations to a single exon can occur because masking or mismatches within exons can split what should be one long matched segment into two or more smaller alignments. In addition, BLAT can generate more than one localization to the exact same region of the genome, as illustrated in Figure 3C. This is a known idiosyncrasy of the BLAT program, and is resolved at the UCSC genome browser Web site by removing such repeat matches [21]. This problem results in no appreciable increase in exon or gene recall compared to SSAHA and MegaBLAST, and also no great loss in precision, as most duplications appear to provide a correct localization (Figure 1). Although we cannot ascertain with certainty how many exons and partial exons each sequence tag spans, we expect that they too generate multiple localizations to a single exon.

Localization to the correct nucleotide

As expected, the greatest variation in the localization results reported by the three programs is at the nucleotide level (Figure 1C). Recall is diminished for SSAHA and MegaBLAST, but remains high for BLAT. SSAHA detects 77% of control nucleotides for gene queries and 7% for sequence tag queries, MegaBLAST detects 89% of control nucleotides for gene queries and 9% for sequence tag queries, and BLAT localizations detect 93% of control nucleotides for gene queries and 9% for sequence tag queries (Figure 1C). Again, recall for sequence tags is so low only because these represent short fragments of genes and so do not contain sufficient information to allow matching a large proportion of the nucleotides comprising the cognate genes.

Diminished recall at the level of individual nucleotides reflects several types of problems, including failure to match to very short exons, misalignment over gaps, and errors in either the query or the genome sequence. The principal cause, however, is difficulty in aligning sequence, using either genes or sequence tags as queries, at the edges of exons. Although failure to accurately align a query to genomic sequence at the edges of exons only slightly lowers the recall levels for each program, each of the three algorithms compared in this study exhibits characteristic problems in localization at the edges of exons, as illustrated in Figure 3C. Figure 4 provides a summary of the performance of each algorithm in exactly matching exon boundaries.

SSAHA correctly matches only 6% of exon boundaries, and only 0.5% of exons (98 of 21,464 total exons) are perfectly matched at both exon edges. The reason for this is that the algorithm splits the genome into non-overlapping fragments that may or may not correlate with exon boundaries. If the edge of an exon does not overlap with an indexed fragment of the genome with a length sufficient to meet the threshold for reporting a match, that fragment will not be included in the match that is returned. Thus, in Figure 3C, SSAHA fails to align 9 nucleotides of both the full-length gene and the sequence tag BG-RRR265 at the 3' edge of the exon because the match does not meet the minimum length of 10 nucleotides. Similarly, small gaps or mismatches that often occur at the ends of sequence tags can interrupt a match, resulting in a minimum loss of 10 nucleotides in the match alignment. (The developers of SSAHA have implemented a new version, SSAHA2 [22], which combines the original SSAHA searching algorithm

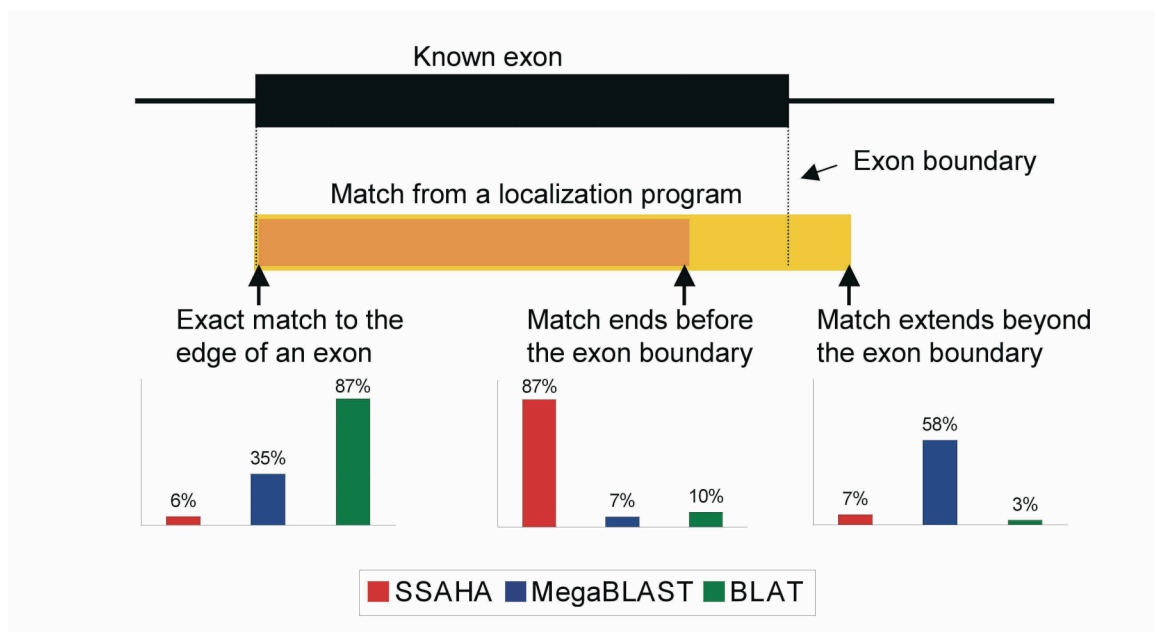


Figure 4.4. A summary of the alignments by each program to the edges of exons.

A representation of an exon is shown at top, with a representation of the three possible match outcomes below, i.e., an exact match to the exon boundary, a match that ends before the exon boundary, and a match that extends beyond the exon boundary. The percentage of all matches by each program that fall into those categories are depicted as bar graphs. Left: Percentage of matches correctly aligned to either exon boundary. Middle and right: Percentage of matches incorrectly aligned to an exon boundary, with the match ending before or extending beyond a boundary, respectively.

with a more sensitive alignment program. The changes incorporated in the new version make it likely that SSAHA2 will behave differently from SSAHA. Additionally, associated programs, such as ssahaEST, combine the search and alignment stages of SSAHA2 with several splice site models to increase detection of exon boundaries. SSAHA2 and its associated programs were not included in this analysis as benchmarking and full documentation has not yet been published, although binaries are now available for download from Ensembl.)

In contrast to SSAHA, MegaBLAST often extends alignments beyond the edges of exons. MegaBLAST localizations align up to, but not beyond, exon boundaries in 35% of attempts, with only 11% of exons receiving perfect alignments at both edges. Moreover, the algorithm generates the longest alignments possible, making no attempt to ensure that each nucleotide in the query sequence is matched only once. Thus, MegaBLAST may extend a match beyond the edge of an exon whenever the adjacent intronic sequence coincidentally matches the query sequence (Figure 3C).

BLAT localizations are the most likely to correctly match exon edges, due to the extra steps the algorithm takes to compute correct exon splice sites and match each nucleotide in the query sequence only once. BLAT localizations match exon edges in 87% of attempts, with 79% of exons perfectly aligned at both edges. It is possible that these rates are slightly inflated by counting multiple overlapping correct localizations that occurred in our automated analysis (see Figure 3C for an example). Even so, BLAT has a clear advantage over SSAHA and MegaBLAST in regard to correct identification of exon boundaries.

With respect to precision at the nucleotide level, SSAHA performs the best, achieving the correct localization 92% of the time for both genes and sequence tags. Precision for MegaBLAST and BLAT is also high, with 85% of both sequence tag and full-length gene localizations matching control nucleotides.

Algorithm speed

Computation times were collected for each localization run. All sequence tags or full-length genes were passed to the localization program as a single file in Fasta format [23] (Table 1). Localizations performed with the unmasked genome were not used in the preceding analysis as this had generally only a small negative effect on recall, but had a large negative impact on precision and analysis speed. (See Figure 3B and associated text for an exception.) SSAHA was the fastest program by about five-fold. MegaBLAST and BLAT were comparable in speed, with BLAT showing an advantage in aligning longer sequences, and MegaBLAST performing more quickly with shorter sequences. Not shown in the Table is the time required for genome indexing, required by both SSAHA and BLAT. This step requires 895.5 seconds for SSAHA, and 399.1 seconds for BLAT, but needs only be run once per genome build.

Conclusions

Overall, analysis of stand-alone versions of the three localization algorithms, SSAHA, MegaBLAST and BLAT, show that all perform well in localizing both full-length genes and sequence tags to the mouse genome. The differences in performance for

sequence tags and full-length reference sequences were surprisingly small, with no program exhibiting significantly diminished performance with sequence tags, despite their generally low quality when compared with full-length reference sequences. While recall and precision performance differ minimally among the programs at the level of gene and exon localization, at a more detailed level, and focusing especially on nucleotide recall, greater variations are found, with different types of characteristic errors associated with each program. Therefore, the choice of the appropriate localization program depends on the specific purpose of the researcher.

As localization to a general region of the genome is performed equally well by all three programs, considerations such as the ease of use of the program and computational speed may become important considerations in choosing which program to use. SSAHA is the fastest program and has the simplest output, so it would seem to be a natural choice for localizing large data sets for general purposes. For automated applications requiring correct localization at the nucleotide level, such as SNP detection or evaluation of alternative splicing, BLAT is currently the best option, as it is distinctly better at aligning the edges of exons. Additionally, the process by which BLAT groups together proximal matches improves the separation between the scores of correct and incorrect matches, increasing confidence in the result. These advantages come at a cost of speed, with BLAT being significantly slower than SSAHA, though comparable in speed with MegaBLAST. For our purpose of localizing gene trap sequence tags to the mouse genome, BLAT was chosen as the program to incorporate into our local annotation pipeline, although use of multiple programs may eventually be implemented to ensure the highest levels of recall and precision.

Methods

Sequences

A set of sequence tags for which the localization of the full-length genes are known was used in this study. The sequence tags were derived from knockout experiments performed by members of the IGTC. Initially, 34,138 sequence tags were annotated by using BLAST to search the GenBank non-redundant database [24] for matching gene transcripts. Only those sequence tags that matched a single transcript with at least 95% identity over a contiguous region of at least 90% of the length of the sequence tag, or matched at least 60 contiguous bases at the 3' end of the sequence tag were considered in our analysis. This eliminated the shortest sequence tags, those that matched with multiple genes or genes with multiple differing transcripts, those that matched genes not yet contained in the GenBank non-redundant database, those that did not match a gene, and all sequence tags generated by trapping processes designed to capture introns rather than exons. Additionally, sequence tags were filtered by requiring that their associated gene transcripts be present at each of the major mouse genome browsers, i.e., Ensembl, NCBI, and UCSC. After filtering, a total of 7,043 sequence tags and 3,320 associated gene transcripts remained [see “Sequence tags” and “Genes” files available from *BMC Genomics* in association with this manuscript]. Half of the localizations were not consistent between all genome browsers, leaving a set of 3369 sequence tags associated with 1659 genes all assigned exactly the same coordinates on the mouse genome. The sequence tags range from 32 to 1023 nucleotides in length (mean 255, median 202) and the genes range from 290 to 64,931 nucleotides in length (mean 3611, median 2485).

Sequence tags and their cognate full-length genes were localized in NCBI Mouse Genome Build 34, the fourth major genome build for the mouse [19]. Build 34 is a composite of high quality high-throughput genome sequence and whole-genome shotgun sequence. Localizations were performed with both an unmasked version of the genome and a version with repeat and low-complexity regions removed by RepeatMasker database version 20050112 [17], which uses RepBase update 9.11 [25]. Except as indicated, the results described below were obtained by searching the masked version of the genome, which is the default practice.

Computation

Alignments were performed on a Hewlett-Packard (HP) AlphaServer GS1280 system, using a single 1.15 GHz processor.

Local versions of online algorithms BLAT, MegaBLAST, and SSAHA were obtained from the genome browser web sites at UCSC, NCBI, and Ensembl, respectively. The most recent versions were chosen, with the exception of BLAT version 26 (February 2004), which was selected because it is the version used to localize BayGenomics sequence tags. SSAHA version 3.1 and MegaBLAST version 2.2.10 represent the most current releases available on July 2005. To approximate the online localization process, parameters were set to match the default parameters employed by the online programs. The three programs do not share the same types of parameters, however, and where the parameters are the same or similar, the values assigned to them are not necessarily consistent. Of particular importance in this study is the default “word length”, i.e., the length of indexed genome fragments. A decrease in word length increases the capacity to

detect short but real matches, but also increases the number of erroneous matches. Word length was set to 10 nucleotides for SSAHA, and 11 nucleotides for BLAT, with a minimum of two contiguous “words” required to seed a match. Similarly, MegaBLAST requires a minimum of 28 contiguous matches to generate an alignment. How each algorithm deals with repeat masking is also important. None of the algorithms seed alignments in masked regions, but BLAT and MegaBLAST can be set to allow alignments to be extended into regions masked by the RepeatMasker algorithm. By default BLAT is set to allow such alignment extensions, but MegaBLAST is not, resulting in the type of differences between alignments presented in Figure 3B. [see Table 2 for a full list of the parameters used for each program.]

A comparison algorithm was devised to demonstrate the accuracy of the localization programs at three levels of granularity relevant for biological inquiry: gene, exon, and nucleotide. At the genic level, any overlap between a localization reported by a program and a known coordinate for a gene was considered a true positive, even if the overlap consisted of a single nucleotide. Similarly, for each exon, only a single nucleotide match was required for a true positive. At the nucleotide level, only an exact match at a single nucleotide position was counted as a true positive. Thus, each level of granularity imposes a different stringency in this analysis. Results are represented by recall and precision scores for each algorithm.

MegaBLAST parameters

-d Database = goldenpath_ucsc_mouse_masked.fa (mouse genome build 34, Fasta file)
-e Maximum allowed expectation value = 1000000.0 (actual maximum was 0.003)
-m alignment view = tabular
-F Filter query sequence = False
-X X dropoff value for gapped alignment (in bits) = 20
-I Show GI's in deflines = False
-q Penalty for a nucleotide mismatch = -3
-r Reward for a nucleotide match = 1
-v Number of database sequences to show one-line descriptions for = 500
-b Number of database sequence to show alignments for = 0
-D Type of output = tab-delimited one line format
-a Number of processors to use = 1
-M Maximal total length of queries for a single search = 20000000
-W Word size (length of best perfect match) = 28
-z Effective length of the database (use zero for the real size) = 0
-P Maximal number of positions for a hash value (set to 0 to ignore) = 0
-S Query strands to search against database = both
-T Produce HTML output = False
-G Cost to open a gap (zero invokes default behavior) = 0
-E Cost to extend a gap (zero invokes default behavior) = 0
-s Minimal hit score to report (0 for default behavior) = 0
-f Show full IDs in the output (default - only GIs or accessions) = False
-U Use lower case filtering of FASTA sequence = False
-R Report the log information at the end of output = False
-p Identity percentage cut-off = 0
-A Multiple Hits window size = 0
-y X dropoff value for ungapped extension = 10
-Z X dropoff value for dynamic programming gapped extension = 50
-t Length of a discontinuous word template (contiguous word if 0) = 0
-g Generate words for every base of the database (default is every 4th base) = False
-n Use non-greedy (dynamic programming) extension for affine gap scores = False
-N Type of a discontinuous word template = coding
-H Maximal number of HSPs to save per database sequence = unlimited

SSAHA parameters

-queryFormat = fasta file
-subjectFormat = fasta file: goldenpath_ucsc_mouse_masked.fa (mouse genome build 34, Fasta file)
-queryType = DNA
-subjectType = DNA
-parserFriendly = pf Show one match per line as a set of tab delimited fields:
 match direction: F forward, R reverse
 query name
 query start
 query end
 subject name
 subject start
 subject end
 number of matching bases

percentage identity

-logMode -lm Controls the output of log information
cerr - send to standard error

-packHits -ph Store position of each word in a "packed"
format comprising 32 bits per word. This halves
the size of the .body file at the expense of a
slight decrease in search speed.

-wordLength = 10

-maxGap = 0 Maximum gap allowed between successive hits for
them to count as part of the same match.

-maxInsert = 0 Maximum number of insertions/deletions allowed
between successive hits for them to count as part
of the same match.

-maxStore = 10000 Largest number of times that a word may occur in
the hash table for it to be used for matching
expressed as a multiple of the number of
occurrences per word that would be expected
for a random database of the same size as the
subject database.

-numRepeats = 0 Maximum size of tandem repeating motif that can be
detected in the query sequence. This option may
produce faster and better matches when dealing
with data containing tandem repeats.

-minPrint = 1 The minimum number of matching bases or residues
that must be found in the query and subject
sequences before they are considered as a match
and thus printed.

-queryStart = 1 Specifies the number of the first query sequence to
be matched with the subject sequences (numbering of
both the query and subject sequences starts at 1).

-queryEnd = not specified Specifies the number of the last query sequence to
be matched with the subject sequences. If not
specified, continues until the end of the query
sequence data is reached.

-reverseQuery = yes When matching the reverse strand of a query,
convert the positions of any matches found
into the coordinate frame of the forward strand.
Has no effect if queryType is set to protein.

-sortMatches = 0 Output only the top n matches for each query,
sorted by number of matching bases, then by
subject name, then by start position in the
query sequence.
Default value is zero, which outputs all matches
for each query and does no sorting.

-stepLength = 10 Number of base pairs gap between words used to
produce hash table. Ignored if a precomputed
hash table is being used. Default value is
equal to wordLength.

-queryReplace = default Specifies behaviour upon encountering unexpected
alphanumeric characters in query sequences:
Default: replace with 'A' for DNA, 'X' for protein

-subjectReplace = tag Specifies behaviour upon encountering unexpected
alphanumeric characters in subject sequences:
tag - 'tag' the word so that it is not put
into the hash table.

-substituteWords = no Look for single base/amino mismatches in words

<p>that occur less than this many times more often than would be expected for a random database of the same size as the subject database.</p> <p>-bandExtension = 0 Specify size of the band to use for banded dynamic programming, when producing a graphical alignment. 0 - diagonal only</p>
<p>BLAT parameters</p> <p>-t Database type = dna: goldenpath_ucsc_mouse_masked.fa (mouse genome build 34, Fasta file) -q Query type = dna - DNA sequence -ooc Use overused tile file = 11.ooc -tileSize sets the size of match that triggers an alignment = 11 -oneOff Mismatches allowed in tile = 0 -minMatch sets the number of tile matches = 2 -minScore This is twice the matches minus the mismatches minus some sort of gap penalty = 30 -minIdentity Sets minimum sequence identity (in percent) = 90 -maxGap sets the size of maximum gap between tiles in a clump = 2 -repMatch sets the number of repetitions of a tile allowed before it is marked as overused = 1024 -minRepDivergence minimum percent divergence of repeats to allow them to be unmasked = 15 -out output file format = psl (Tab separated format without actual sequence)</p>

Table 4.2. Parameters for each program used. Descriptions of the parameters have been adapted from the accompanying documentation for the MegaBLAST, SSAHA, and BLAT stand-alone programs.

Author's Contributions

CAH constructed the sequence sets, analyzed the localization data, devised the comparison method and drafted the manuscript. CCH, TEF, and PCB participated in the design of the study. PCB helped to draft the manuscript and conceived of the study. DS installed the localization algorithms, performed the localizations and participated in the design of the study. MK constructed initial test sequence sets. All authors contributed to and approved the final manuscript

Acknowledgments

This work was supported by NIH grants U01 HL66600 and P41 RR01081, and by an NSF fellowship award to CAH.

References

1. Stanford WL, Cohn JB, Cordes SP: **Gene-trap mutagenesis: past, present and beyond.** *Nat Rev Genet* 2001, **2**(10):756-768.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
3. Stryke D, Kawamoto M, Huang CC, Johns SJ, King LA, Harper CA, Meng EC, Lee RE, Yee A, L'Italien L *et al*: **BayGenomics: a resource of insertional mutations in mouse embryonic stem cells.** *Nucleic Acids Res* 2003, **31**(1):278-281.
4. Nord AS, Chang PJ, Conklin BR, Cox AV, Harper CA, Hicks GG, Huang CC, Johns SJ, Kawamoto M, Liu S *et al*: **The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse.** *Nucleic Acids Res* 2006, **34**(Database issue):D642-648.
5. Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, Young SG, Ruiz P, Soriano P, Tessier-Lavigne M *et al*: **A public gene trap resource for mouse functional genomics.** *Nat Genet* 2004, **36**(6):543-544.
6. Townley DJ, Avery BJ, Rosen B, Skarnes WC: **Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice.** *Genome Res* 1997, **7**(3):293-298.
7. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**(3):175-185.
8. Chen D, Patton JT: **Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension.** *Biotechniques* 2001, **30**(3):574-580, 582.
9. Wiles MV, Vauti F, Otte J, Fuchtbauer EM, Ruiz P, Fuchtbauer A, Arnold HH, Lehrach H, Metz T, von Melchner H *et al*: **Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells.** *Nat Genet* 2000, **24**(1):13-14.
10. Zhang Z, Schwartz S, Wagner L, Miller W: **A greedy algorithm for aligning DNA sequences.** *J Comput Biol* 2000, **7**(1-2):203-214.
11. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases.** *Genome Res* 2001, **11**(10):1725-1729.
12. Kent WJ: **BLAT--the BLAST-like alignment tool.** *Genome Res* 2002, **12**(4):656-664.
13. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29**(1):11-16.
14. Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV: **The Ensembl Web site: mechanics of a genome browser.** *Genome Res* 2004, **14**(5):951-955.
15. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ *et al*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31**(1):51-54.

16. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6**(1):31.
17. **RepeatMasker Open-3.0.** [<http://www.repeatmasker.org>]
18. Mighell AJ, Smith NR, Robinson PA, Markham AF: **Vertebrate pseudogenes.** *FEBS Lett* 2000, **468**(2-3):109-114.
19. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P *et al*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**(6915):520-562.
20. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ *et al*: **The DNA sequence of human chromosome 22.** *Nature* 1999, **402**(6761):489-495.
21. **pslReps selects the best alignments for a particular query sequence, using a 'near best in genome' approach.**
[<http://genome.ucsc.edu/goldenPath/help/blatSpec.html>]
22. **SSAHA2** [<http://www.sanger.ac.uk/Software/analysis/SSAHA2>]
23. Pearson WR: **Rapid and sensitive sequence comparison with FASTP and FASTA.** *Methods Enzymol* 1990, **183**:63-98.
24. Benson DA, Boguski M, Lipman DJ, Ostell J: **GenBank.** *Nucleic Acids Res* 1996, **24**(1):1-5.
25. Jurka J: **Rebase update: a database and an electronic journal of repetitive elements.** *Trends Genet* 2000, **16**(9):418-420.

Analysis of Alternative Splicing Utilizing Microarray Experiments

Introduction

Alternative splicing research was in its infancy in 2001, when the research detailed here was performed. At that time, only a few large-scale studies of the prevalence of splice variation had been performed, and most mammalian genomes had not been sequenced, although complete sequences had recently been made available for the human and mouse genomes. Similarly, microarray expression experiments and high-throughput proteomics methods were recent additions to the genomic research toolbox, and had not been perfected to the extent that they have been today. The novelty of the field of alternative splicing research provided great opportunities to develop new methods to detect splice variation with these exciting new tools.

Background

Alternative splicing occurs during the intron removal process when an exon is skipped, inserted, or spliced at a different site than in the reference mRNA. It allows multiple transcripts to be generated from the same gene, which increases the potential range of proteins that can be produced by a genome. Alternative splicing has also been found to be a regulatory mechanism in certain eukaryotic cell growth processes and to

play a role in tissue differentiation [1]. Alternative splicing is also known to affect the working of drugs and to play a role in some human diseases [2]. A few simple examples of alternative splicing are shown in Figure 1.

Alternative splicing is thought to be a primary mechanism by which cells produce a more varied set of proteins than they have genes. This hypothesis is supported by evidence that a large percentage of human genes give rise to multiple mRNA transcripts. In 2001, estimates of the percentage of human genes that undergo alternative splicing at the mRNA level ranged from 35% [3, 4] to 55% [5]. Today, the percentage is estimated to be greater than 60% [6]. Other eukaryotes show prevalent alternative splicing, with rates in mouse [1] and rat approaching that of humans. Some alternative splice sites are conserved between species [6].

When I began this research, the only method in use for systematic detection of splice variants was the alignment of expressed sequence tags (ESTs) to full-length mRNAs, or to genomic sequence. Splice variants are apparent in alignments with unmatched sequence flanked by multiply aligned sequence. This remains the primary method used today. Several databases house information about splice variants that has been generated in this manner. The Putative Alternative Splice Database (PALS) [7] contained over 14,000 human genes and 8,000 mouse genes in which alternative splicing had been detected in 2001, approximately half the genes it contains today. The Alternative Splicing Database (ASDB) [8] contained splice variants for six different organisms in 2001, and 181 presently. The Human Alternative Splicing Database [9] contained over 6,000 splice variants in 2001, and has almost 22,000 now.

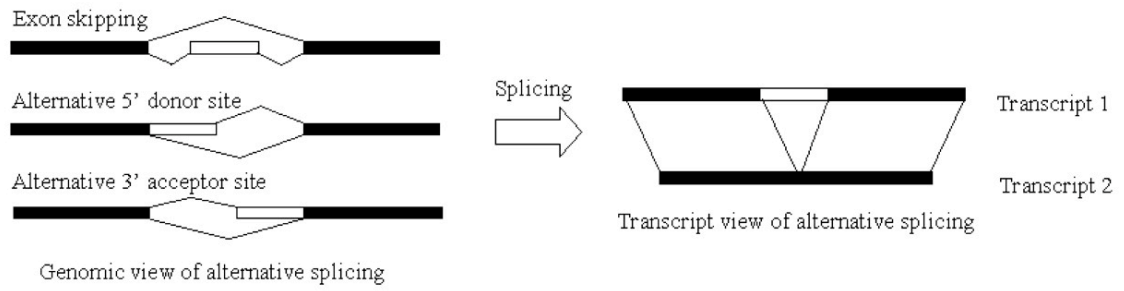


Figure 5.1. The three types of alternative splicing analyzed in this research. Adapted from Huang, et al. 2002 [7].

Detection of splice variation using EST sequences has a few drawbacks. Databases containing ESTs are not complete, and are not guaranteed to have a set of sequences that represent transcripts of all genes. ESTs are also of generally lower sequence quality than full-length mRNA sequences, although the quality has been improving as sequencing technology becomes more advanced. This can create problems with the alignments used to predict the presence of splice variants. Finally, alignments of ESTs to mRNA sequences can indicate the existence of a splice variant, but not the prevalence of that transcript or its variance in different environmental or developmental conditions.

A method with the potential to overcome some of these problems is the use of microarray expression data to detect alternative splicing. As illustrated in Figure 2, in gene expression experiments, RNA is extracted from a tissue sample, labeled with a fluorescent marker, and washed over a microarray. Microarrays are chips carrying hundreds or thousands of probes that are complementary to the sequence of a gene of interest, most often with several probes to represent each gene. The probes hybridize with the RNA sequences, and retain fluorescent signal after the excess labeled-RNA is washed away. This signal is used to show the presence and prevalence of a known set of genes. As splice variant transcripts have missing or alternative sequence as compared with canonical gene sequences, probes designed to detect these differences may be able to indicate occurrences of alternative splicing.

The most unambiguous way to determine whether alternative splicing can be detected by comparing differences in fluorescent signal, or hybridization value, between probes from a single gene would involve testing a tissue sample in which the splice

Microarray Overview

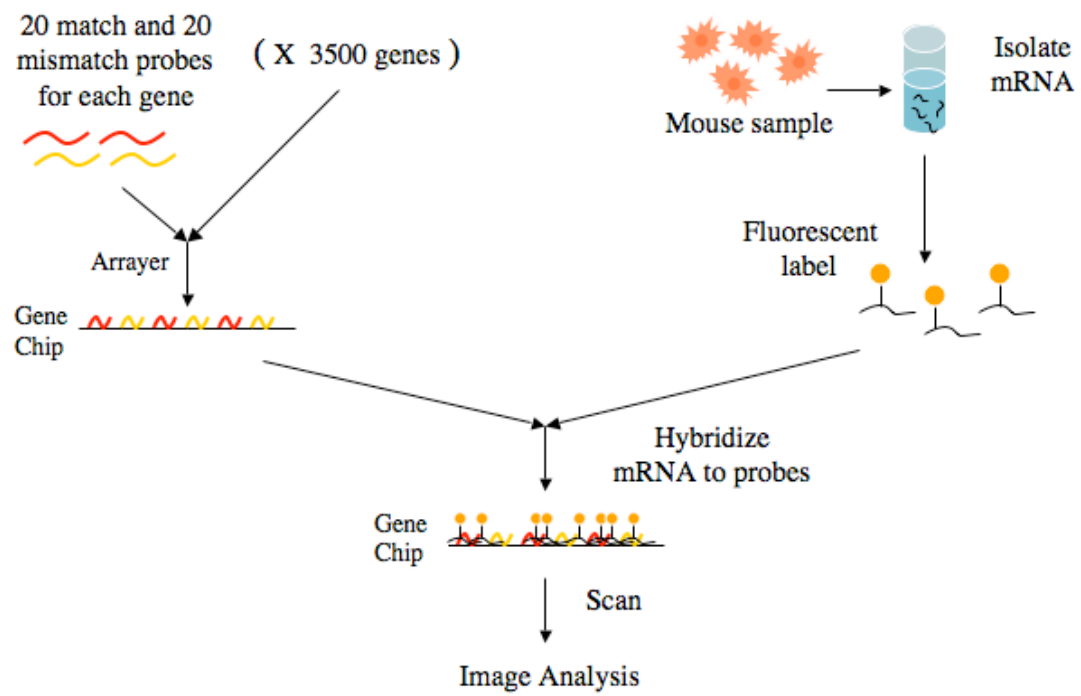


Figure 5.2. Overview of the microarray experimental process.

variants present were known. One way to approach this would be to create an EST library for a tissue, use that library to find a set of genes that produce splice variants, produce a microarray carrying splice variant-specific probes for these genes, hybridize RNA from the original tissue sample to this microarray, and find what proportion of splice variants detected by EST alignment are also detected with the expression data. This approach was feasible in 2001, but would have been expensive, as microarray printing was not as prevalent or economical as it is today.

A faster and free approach would be to approximate known splice variants in a tissue with all known splice variants for an organism and utilize pre-existing expression data that contains genes associated with these splice variants. While it would not be expected that all splice variants are expressed at all times, using data from a variety of test samples in different environments would increase the likelihood of capturing conditions under which at least some splice variants are present at detectable levels. Similarly, whereas microarrays designed specifically for an experiment are ideal, the practice of spreading probes over an extended region of a gene suggests that even microarrays not designed to detect splice variants will contain probes that align to regions of a gene that are present in only some variants.

The magnitude of the fluorescent signal for each probe depends on the number of RNA molecules that hybridize with that sequence. Therefore, in genes that produce multiple mRNA transcripts, some of which are missing a fragment of sequence it would be expected that hybridization values would vary across the length of the gene and would be lessened in regions not present in all transcripts. Certain types of splice variation result in such transcripts. In particular, exon skipping (Figure 1), which is the most

common type of alternative splicing, results in transcripts missing one or more exons. It should be possible to detect splice variants by investigating clusters of probes with lower hybridization values than probes situated elsewhere in a gene, as shown in Figure 3. An advantage of this method would be wide applicability to many expression data sets, which could be re-analyzed for new findings vis-à-vis alternative splicing.

The use of microarray expression data is limited by the fact that genes must be known in order to design probes that will hybridize to their transcribed mRNA sequences. However, even in the early days of microarray design, scientists foresaw the possibility of detecting novel genes by designing microarrays with probes designed over the length of entire genomes [10]. Transcription of a novel gene would lead to mRNA with the potential to bind any probes located in the coding regions of that gene. Similarly, these probes might show variability in hybridization values as a result of splice variation that affects the makeup of mRNA produced by a gene.

Methods

This analysis required a large and diverse set of expression data, the set of known splice variants associated with the genes represented on the microarrays, and a method for determining positive and negative results.

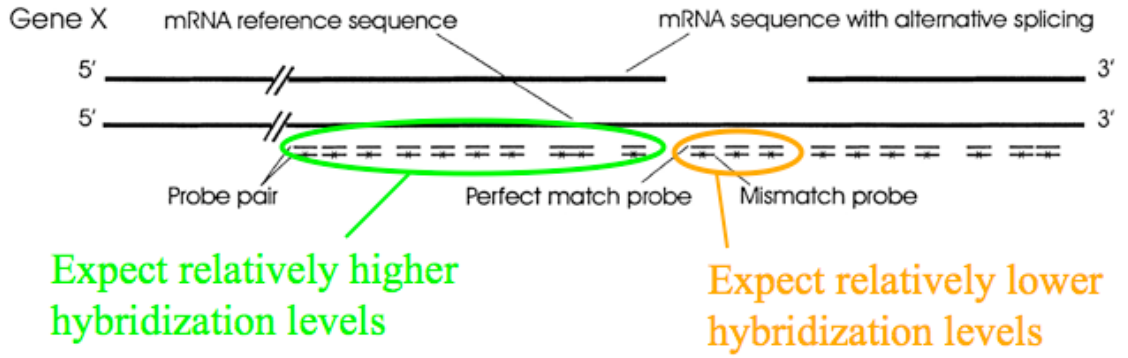


Figure 5.3. Graphical representation of perfect match and mismatch probes aligned to two transcripts, the first a splice variant missing an exon, and the second the full-length transcript of a gene. Lower hybridization values would be expected from probes that hybridize to regions of a gene that are not present in all transcripts. (Adapted from Hu, et al. 2001 [11]).

Expression Data Set

In order to increase the chances of detecting splice variants, I chose one of the most comprehensive sets of microarrays available, the mouse Mu6500 Gene Chips developed by Affymetrix (Santa Clara, CA). This set of four microarrays represented nearly all of the mouse genes known at that time, as well as a great number of EST sequences thought to represent genes that had yet to be fully sequenced.

Prof. Bruce Conklin, a frequent collaborator with the Babbitt Lab, used the Mu6500 Gene Chips to assess expression differences in mouse tissue as a result of cardiovascular insult. One large data set, created by Dr. Kam Dahlquist, contained expression data from each of the four Mu6500 Gene Chips for 30 mice representing one of four experimental conditions: control mouse heart tissue, heart samples from mice after two weeks of increased G protein expression, eight weeks of increased G protein expression, and mice in recovery [12]. This data is available from the PNAS web site (<http://www.pnas.org/>). This data set was ideal for my purposes, as it contained not only data from wild-type mice, but would also allow for the detection of changes in the prevalence of splice variants due to experimental conditions, should the detection method prove successful.

In total, the Mu6500 Gene Chips contain 6,519 sets of probe sequences. In addition to the series of 200 control probes among the four chips, there are 6,319 “full sequences” listed by GenBank accession, 3,281 of which were linked to known genes and 3,038 of which were linked to ESTs. Since ESTs can be of poor sequence quality and contamination from other species is a known problem, a filtering step was taken to ensure

that the “full sequences” were actually mouse transcripts. In order to be considered for this study, the GenBank accession was required to be present in a Unigene cluster [13], which is multiple alignment of mRNA and EST sequences that indicate similarity presumed to stem from the presence of a single gene sequence. Of the 6,319 GenBank accessions, 4,651 were found in Unigene clusters.

In performing this filtering step, it became apparent that duplicated and out-of-date sequences were represented on the MU6500 Gene Chips. Due to the continual merging by GenBank of accessions representing identical sequences, accessions for 126 “full sequences” had been replaced with accessions already represented on a microarray. Additionally, 643 accessions mapped to Unigene clusters already present in the dataset, indicating that sequences with different GenBank accessions but corresponding to the same gene were mistakenly considered unique genes in creating the microarrays. Furthermore, GenBank accessions for 388 of the genes were listed as ‘withdrawn’ at the GenBank web site, and were thus excluded from analysis. These problems were not unexpected, since many mouse genes had yet to be annotated and the mouse genome was in a less complete state at the time of this analysis (July 11, 2001). After removing these sequences from consideration, 3,494 accessions remained for splice variant detection.

Association with Splice Variants

Although there are dozens of types of splice variation that can produce different mRNA sequences from a given genic region, only alternative splicing which results in transcripts that are missing an internal fragment of sequence when compared with the

canonical transcript were used in the first pass of this detection method. These types of alternative splicing are shown in Figure 1. Splice variants missing sequence at the 3' or 5' end were not used, as it could not be determined if these were true splice variants or incomplete sequences.

To determine which genes present on the microarrays had known splice variants, I intersected the GenBank accessions associated with the Mu6500 Gene Chips with the GenBank accessions present in the PALS database, which was the largest source of splice variants available at the time. Of the 3,494 GenBank accessions remaining after filtering, 1,327 were present in the PALS database, which contains both mRNA and EST sequences. Information extracted from individual PALS database entries regarding splice variation type was used to filter out sequences not associated with splice variants missing an internal fragment of sequence.

The final filtering steps involved associating the probes themselves with the gene sequences they represented. Not all genes with evidence of the types of splice variants shown in Figure 1 had probes that were located in the appropriate position to detect those variants. Some did not have probes that aligned with exons present in only a portion of transcripts, or only had probes that aligned to those occasionally missing exons. As the point was in intragenic comparison of hybridization values, these genes were not used in the experiment. However, these sequences and their associated information were retained, for they would be interesting to study if a correlation with hybridization values was noted in the principal experiment.

Not all probes associated with the 584 remaining gene sequences were used in the analysis. Approximately 20% of the probes did not align perfectly to both the mouse

genome and the mRNA or EST sequence they were designed to match. The mismatches are presumably due to alterations in the sequences contained in the GenBank database since the production of the Mu6500 Gene Chips and the fact that the mouse genome was not available when these microarrays were designed. These probes were excluded from the analysis, but in all cases sufficient probes remained to retain their associated gene in the analysis. Probes that aligned over a splice junction involved in splice variation were also excluded from the set, although they were kept aside for possible future analysis. All remaining probes were assigned to one of two groups: probes matching conserved regions present in all known transcripts of a gene, and probes matching variant regions which are missing in one or more mRNAs or ESTs. In total, the 584 gene sequences were represented by 4,887 variant region probes and 5,305 conserved region probes, and over the 30 experiments yielded 17,520 hybridization value comparisons. A breakdown of this information by Mu6500 Gene Chip is given in Table 1.

Splice Variant Detection Method

To test for a general effect of alternative splicing on microarray expression experiments, the hybridization values for all variant region probes were compared with the hybridization values for all conserved region probes. Hybridization values were collected using GeneChip 3.1 automated analysis software provided by Affymetrix ([12], supplemental data). Inter-array hybridization values were scaled by setting the total fluorescence intensity of each array, excluding the highest and lowest 2% of readings, to

Chip	Genes	# Variant Region Probes	# Conserved Region Probes	Avg. Variant Region Hyb. Value	Avg. Conserved Region Hyb. Value	Avg. Hyb. Difference
A	149	1203	1267	1240.8	1333.0	+92.2
B	142	1020	1337	1076.8	1122.9	+46.1
C	136	1339	1252	1194.1	1262.5	+68.4
D	157	1325	1449	1062.5	1073.6	+11.1

Table 5.1. A breakdown of the experimental results by Affymetrix MU6500 Gene Chip.

a fixed value. Within each array, hybridization values were normalized by subtracting from the fluorescence intensity of each perfect-match probe the intensity of its paired mismatch sequence.

To identify data that support individual alternative splicing predictions, probe sets for each of the 584 genes in consideration were searched for patterns consistent with known splice variants. A positive result for an individual gene required the hybridization values of no less than 80% of the variant region probes to be lower than the hybridization values of the surrounding conserved region probes. Instances where between 50% and 80% of the hybridization values of variant region probes were lower than the minimum hybridization value for a conserved region probe were considered a negative result, but were put aside for possible detailed analysis in the future.

Results

Alternative splicing appears to have a small general effect on the hybridization values obtained with microarray experiments. As shown in Table 1., differences between the hybridization values of probes aligning to conserved and variant regions of genes were cumulatively positive, indicating that transcripts containing variant region sequence were slightly less prevalent than their associated full-length gene transcripts. However, these differences were small in comparison to the hybridization values themselves, and varied widely by gene and across microarray experiments (Figure 4). Another indication of the relative insignificance of the lower prevalence of variant transcripts is that the

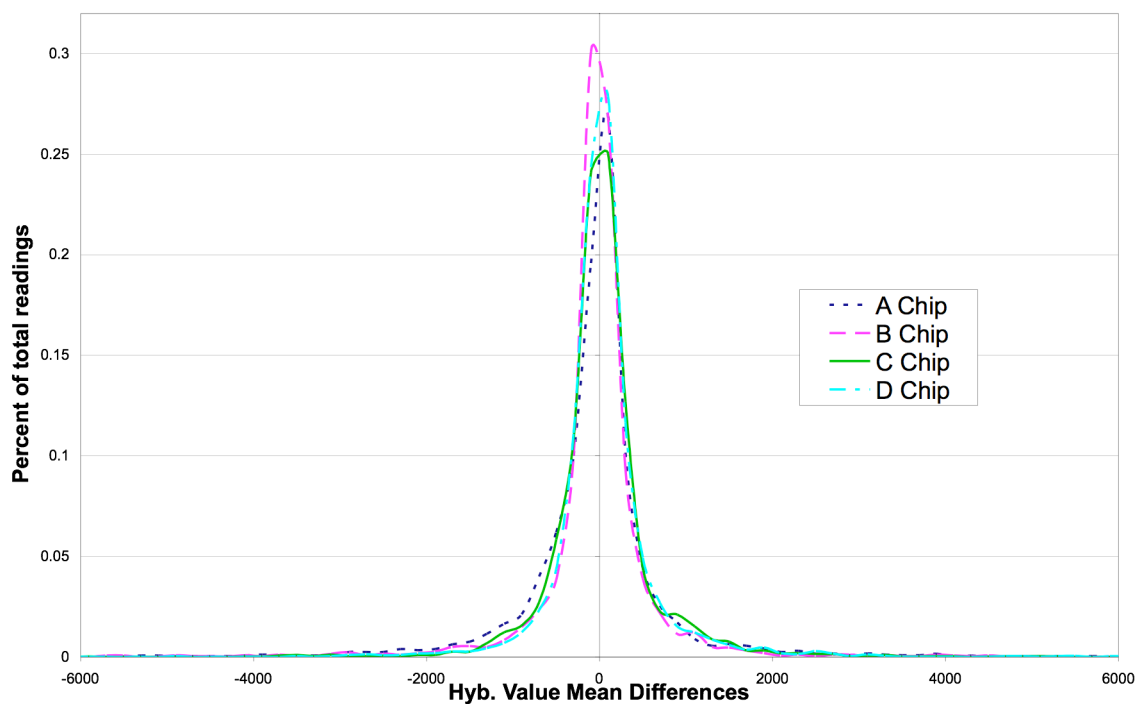


Figure 5.4. Intragenic hybridization value differences between conserved region probes and variant region probes for all probe sets on each of four Mu6500 Gene Chips.

distributions of hybridization values for conserved and variant region probes are not significantly different, as shown in Figure 5.

To investigate whether the small overall effect was due to strong effects from individual genes with prevalent alternative splicing, results for each gene were examined for patterns supporting the presence known splice variants. Probe sets in which >80% of probes in variant regions had lower hybridization values than probes in conserved regions were considered to support the hypothesis that alternative splicing had occurred in that gene in under the associated experimental conditions. Out of 17,520 comparisons, 1,050 showed patterns supporting alternative splicing. These were associated with 93 of the 584 genes examined. This proportion is not out of line with the portion of genes that would be expected to have hybridization patterns consistent with a positive result by chance.

In order to determine whether the results that support alternative splicing were obtained by chance or the presence of splice variants, a control data set was gathered by selecting probe sets in which >80% of probes in conserved regions had lower hybridization values than probes in variant regions. There were 1,423 such probe sets associated with 99 genes. This data is referred to as ‘opposing’ data because not because it negates the presence of splice variants, but because it is taken from randomly distributed data and is thus a control for the hypothesis that the ‘supporting’ data indicates the presence of splice variants in the mouse tissue samples. Differences between the supporting and opposing probe sets would indicate that the supporting data represents more than statistical noise, whereas similarities between the probe sets belie that proposal. Figure 6 gives a graphical representation of the supporting data, opposing

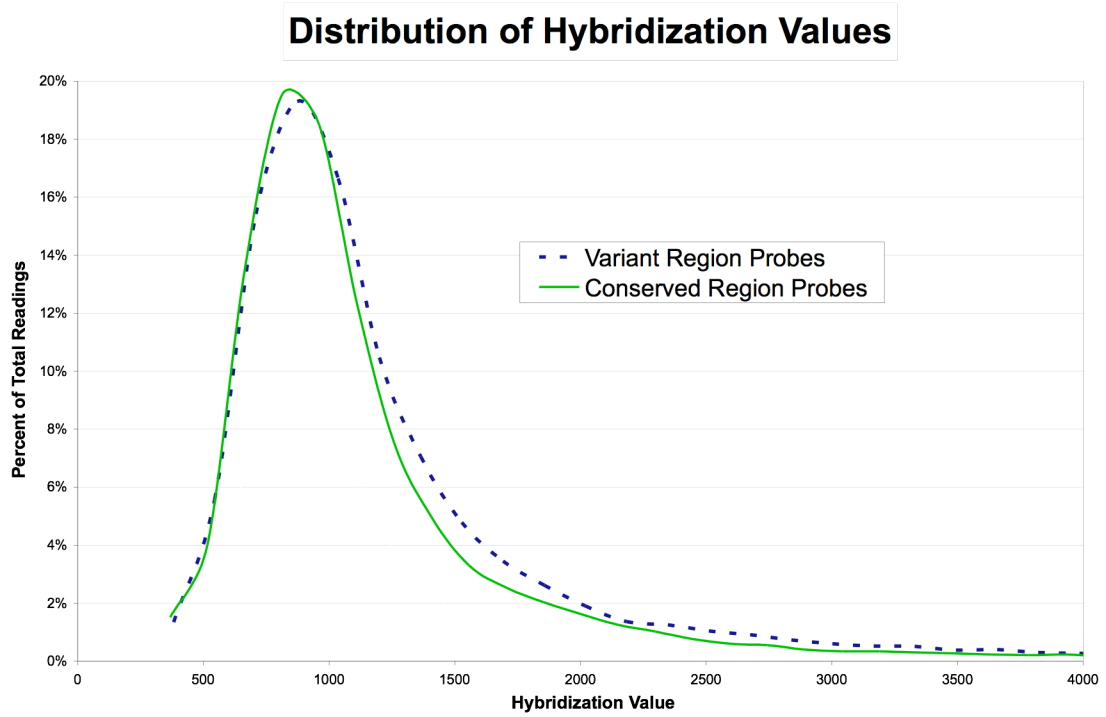


Figure 5.5. Distribution of hybridization values for probes present in all transcripts and probes absent in those transcripts that have been alternatively spliced.

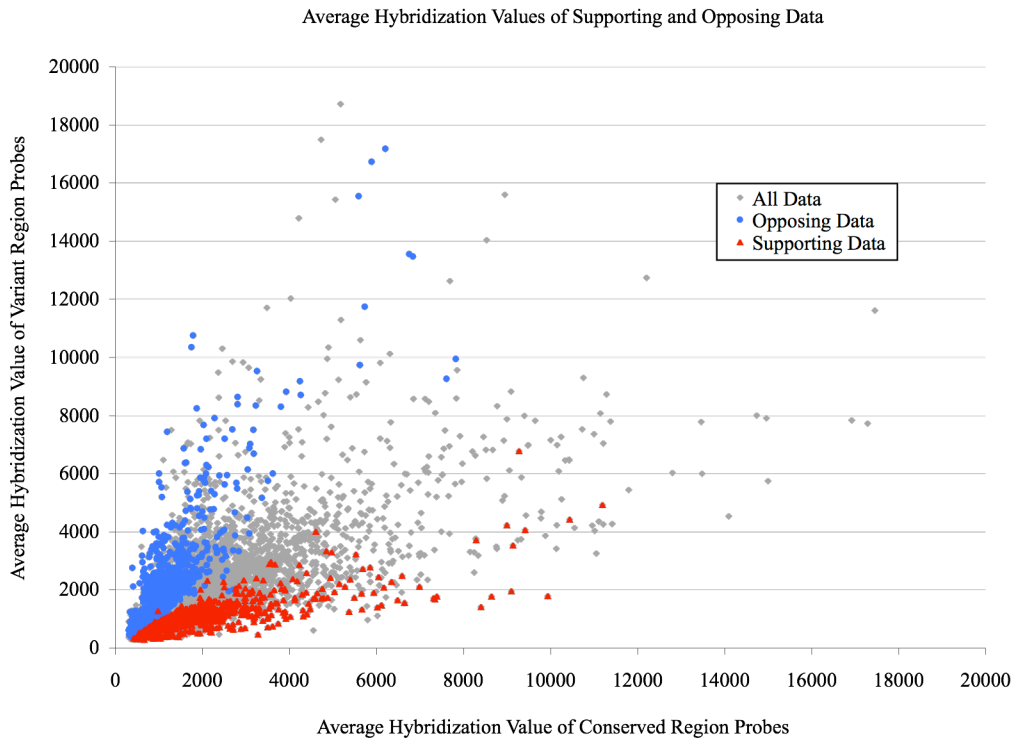


Figure 5.6. Mapping of 17,520 probe sets by the average hybridization value of probes in conserved and variant regions. Probe sets in which >80% of variant region probes have lower hybridization values than probes in conserved regions are indicated in red triangles. Probe sets with the inverse values, in which >80% of conserved region probes have lower hybridization values than probes in variant regions are shown in blue circles.

data, and all 17,520 probe sets by the average hybridization values of their conserved and variant region probes. A comparison of the number of probes in the supporting and opposing probe sets that lie in conserved or variant regions is shown in Figure 7. The cumulative distribution of the differences in average hybridization values for conserved and variant region probes is given in Figure 8. These metrics clearly show that the sets of probes that seem to support the presence of splice variants are not different than the sets of probes with inverse values. Therefore, the small difference in overall hybridization values is best attributed to chance, and patterns consistent with alternative splicing are most likely the result of the random variation common to microarray hybridization values.

Conclusions

Factors that may explain why splice variant transcripts cannot be reliably detected in microarray data involve the quality and quantity of splice variant data available and the true prevalence of splice variant transcripts.

It is not known what portion of alternative splicing is represented in current EST databases. The rapid rise in the number of EST-predicted splice variants since 2001 is an indication that the number is likely to increase along with the size of sequence databases. It is possible that alternative splicing is even more prevalent than currently predicted, and that the probes that are predicted not to be excised in currently known variants are actually missing in splice variants that have not yet been identified.

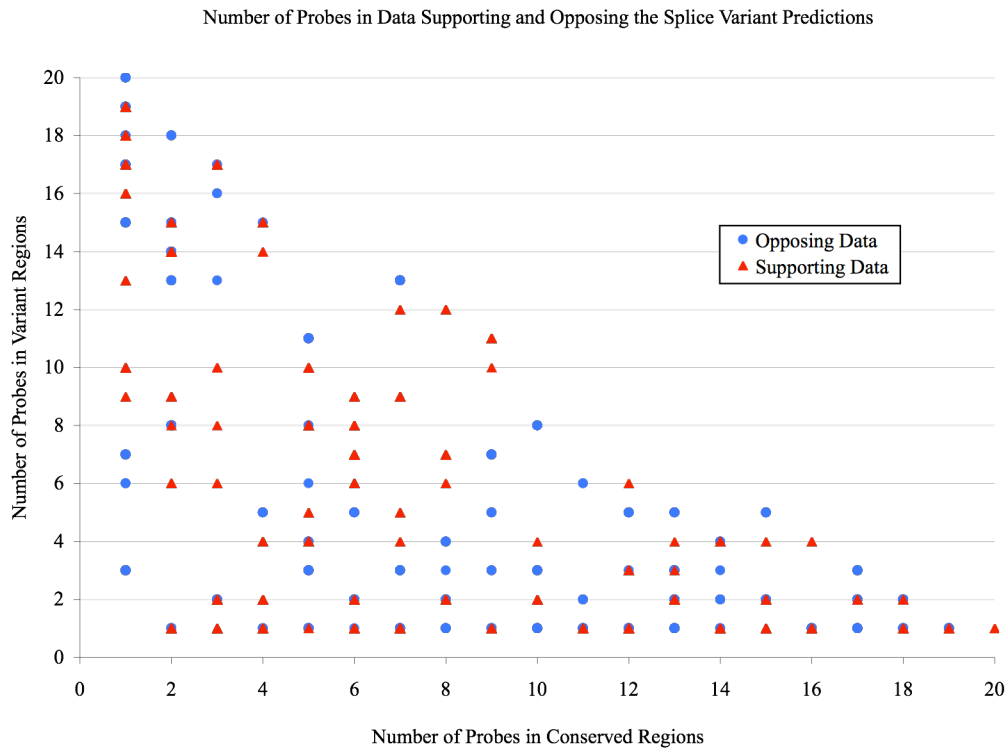


Figure 5.7. The number of probes in conserved and variant regions for each probe set. Probe sets for the 93 genes in which some experiments support the presence of splice variants are shown by red triangles. Probe sets for the 99 genes with values inverse to the supporting data are shown by blue circles.

Distribution of Hybridization Value Differences Between Variant and Conserved Region Probes

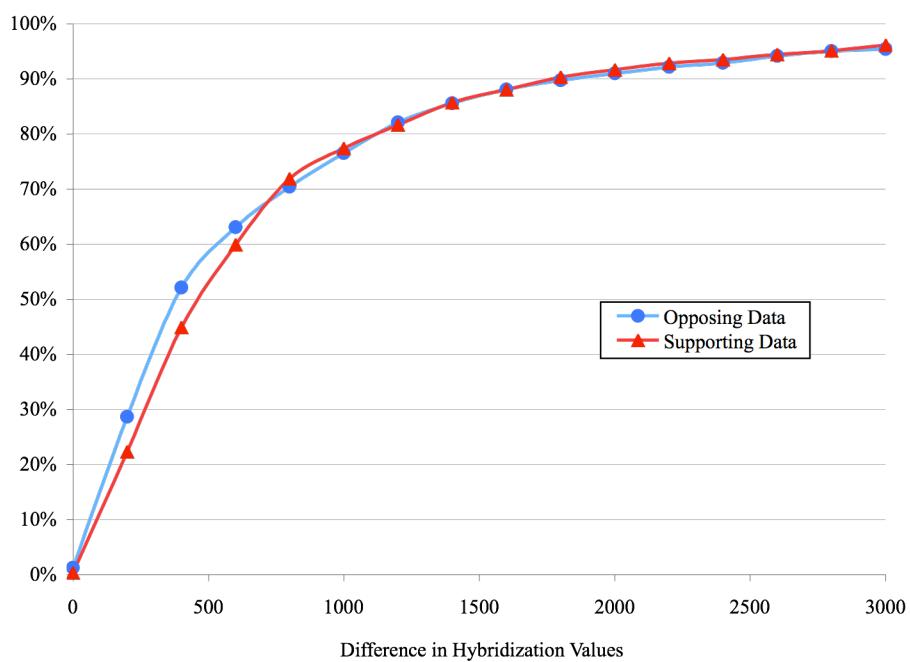


Figure 5.8. Cumulative distribution of hybridization value differences between variant and conserved region probes for opposing and supporting probe sets.

Additionally, alternative splicing supported by only a single EST may be incorrectly predicted, due to chimerism in ESTs, contamination with DNA from other species, and poor EST sequence quality. As the quality and quantity of sequence available continues to improve, the set of known splice variants will become more reliable, and thus more useful in transcript prediction.

It is also possible that alternative splicing represents a very small portion of total mRNA splicing. Microarray data is noisy – possibly at a level that would mask real differences in hybridization levels between spliced out and retained regions. The results detailed in this study might be expected if only a small percentage of transcripts undergo the alternative splicing in question.

More recently, more reliable methods have been developed to utilize microarray experiments to detect alternative splicing. Analysis of hybridization patterns of rat mRNA to microarray probes across a single gene has been used to predict splice variants, three of which were confirmed by sequencing transcripts [11]. Similarly, known alternatively spliced regions can be detected using mRNA microarray chips with splice site-specific probes [14]. Alternative splicing has been detected by polymerase colony technology, whereby solid-phase templates of individual RNA molecules give rise to colonies of amplification products [15]. These methods offer more precision than using pre-existing data to predict splice variation, and are being successfully used to increase the amount of information available in regards to alternative splicing that is available to the scientific community.

References

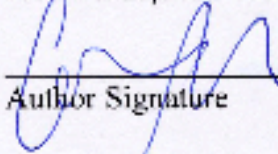
1. Kochiwa H, Suzuki R, Washio T, Saito R, Bono H, Carninci P, Okazaki Y, Miki R, Hayashizaki Y, Tomita M: **Inferring alternative splicing patterns in mouse from a full-length cDNA library and microarray data.** *Genome Res* 2002, **12**(8):1286-1293.
2. Chandrasekharan NV, Dai H, Roos KL, Evanson NK, Tomsik J, Elton TS, Simmons DL: **COX-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: cloning, structure, and expression.** *Proc Natl Acad Sci U S A* 2002, **99**(21):13926-13931.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
4. Mironov AA, Fickett JW, Gelfand MS: **Frequent alternative splicing of human genes.** *Genome Res* 1999, **9**(12):1288-1293.
5. Kan Z, Rouchka EC, Gish WR, States DJ: **Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.** *Genome Res* 2001, **11**(5):889-900.
6. Sharov AA, Dudekula DB, Ko MS: **Genome-wide assembly and analysis of alternative transcripts in mouse.** *Genome Res* 2005, **15**(5):748-754.
7. Huang YH, Chen YT, Lai JJ, Yang ST, Yang UC: **PALS db: Putative Alternative Splicing database.** *Nucleic Acids Res* 2002, **30**(1):186-190.
8. Dralyuk I, Brudno M, Gelfand MS, Zorn M, Dubchak I: **ASDB: database of alternatively spliced genes.** *Nucleic Acids Res* 2000, **28**(1):296-297.
9. Modrek B, Resch A, Grasso C, Lee C: **Genome-wide detection of alternative splicing in expressed sequences of human genes.** *Nucleic Acids Res* 2001, **29**(13):2850-2859.
10. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engle P, McDonagh PD, Loerch PM, Leonardson A, Lum PY, Cavet G *et al*: **Experimental annotation of the human genome using microarray technology.** *Nature* 2001, **409**(6822):922-927.
11. Hu GK, Madore SJ, Moldover B, Jatkoe T, Balaban D, Thomas J, Wang Y: **Predicting splice variant from DNA chip expression data.** *Genome Res* 2001, **11**(7):1237-1245.
12. Redfern CH, Degtyarev MY, Kwa AT, Salomonis N, Cotte N, Nanevicz T, Fidelman N, Desai K, Vranizan K, Lee EK *et al*: **Conditional expression of a Gi-coupled receptor causes ventricular conduction delay and a lethal cardiomyopathy.** *Proc Natl Acad Sci U S A* 2000, **97**(9):4826-4831.
13. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes.** *J Mol Med* 1997, **75**(10):694-698.
14. Clark TA, Sugnet CW, Ares M, Jr.: **Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays.** *Science* 2002, **296**(5569):907-910.
15. Zhu J, Shendure J, Mitra RD, Church GM: **Single molecule profiling of alternative pre-mRNA splicing.** *Science* 2003, **301**(5634):836-838.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

 8/30/07
Author Signature Date