Assessment of Utility of ESTs for Nucleotide Diversity Using Available Assembled Alignments from dbESt, STACK 2.0 and STACK-INDEX

Brian Karlak **Winston Hide**
bkarlak@sanbi.ac.za winhide@sanbi.ac.za
South African National Bioinformatics Institute, University of the Western Cape
Private Bag X17, Bellville, South Africa 7535

1 Introduction

Single Nucleotide Polymorphisms (SNPs) in virtual expressed gene fragment alignments represent a potentially significant resource for both the detection of non-coding and coding, sequence variations. We have clustered and assembled 767 866 human ESTs into 76 131 alignments localised to specific tissues [1]. In addition, we have clustered and aligned 300 000 consensus sequences and unclustered ESTs to generate a comprehensive human gene index of over 38 000 unique linked virtual transcripts (STACK- INDEX) with associated alignments. The resulting dataset is a potentially rich resource for the detection and characterisation of alternate splicing and polymorphisms. Public access to these data will allow investigators to add functional and scientific value to the emerging human gene sequences [2]. We have surveyed the dataset and have developed an initial set of criteria for assessment of possible high likelihood SNPs. We have studied the protein p53 as a model for the system.

2 Detection of SNPs

In order to develop clear understanding of the value of ESTs for SNP analysis, We assayed STACK2.0 alignments for their utility as a resource for detecting SNPs. We developed a semi-automated EST assembly and discrepancy detection system, autoSNP. We then employed this system to establish rules for the detection of SNPs.

3 Implementation

The system sends a seed entry (either EST, STACK2.0 EST consensus alignment or full length mRNA) to dbESt [3] and STACK2.0 to find matching ESTs. If an existing alignment exists it is surveyed, and/or the sequences are assembled using PHRAP (Phil Green, unpublished) and are then processed for discrepancies. The SNP detection procedure is to find "discrepancies" in an alignment of ESTs, and then progressively reject these discrepancies based on the likelihood that they are simply experimental artifacts. "Discrepancies" are simply any positions in an alignment which do not agree. IUPAC codes, uncalled positions, and gaps are ignored. "Agreement" is measured by comparison with the full length mRNA (flmRNA) if available, otherwise the consensus of the alignment is used as a standard.

3.1 Initial Criteria

1) We reject the following:
   a) Discrepancy within 100 bp of end of EST and surrounded by many IUPAC reads, gaps, and other potential discrepancies.
   b) Discrepancy very close / contiguous with another discrepancy.
   c) ESTs that have IUPAC reads throughout the sequence.

2) Library source information can be used to determine if discrepancies are valid. If two ESTs come from the same library but do not contain the same discrepancy, (i.e., one EST has a discrepancy and the other matches the flmRNA or the ESTs have different discrepancies) the discrepancy is rejected.

4 Derived Rules and Results

1) The most effective method for EST detection is to use a full-length mRNA (flmRNA) from Entrez [4] to "fish" for homologous EST sequences.
2) It is important to cull non-human ESTs from the retrieved homologous sequences.
3) BLAST2 [5] is superior to BLAST [6], giving more matches with fewer highest scoring pairs (HSPs) per match. Fewer HSPs per match leads to more accurate probability (P) scores.
4) EST consensii were useful for SNP detection, but short consensus sequences can often simply duplicate information found in other ESTs.
5) EST quality varies immensely. Some ESTs are all high-quality, some have low-quality tails, and some are poor throughout. Poor-quality ESTs are rejected in the assembly & detection steps.
6) STACK2.0 alignments were more comprehensive than STACK 1.0 alignments. It was still preferable to cull the latest dbEST submission for ESTs and to use the STACK consensus as a probe for these ESTs. Consensus sequences proved more useful than ESTs for deriving ESTs from sequences that had no flmRNA.
7) More reliable results were derived from alignments containing several copies of each potential SNP.

MAPPING OF EST DIFFERENCES TO KNOWN P53 MUTATIONS:

We selected the cancer-causing gene p53 as a model for the detection of potential cSNP (coding single nucleotide polymorphism) variations. While each EST has many failed reads ('N'), there are few base substitutions. Fig. 1 shows a partial alignment of the ESTs we analysed vs. flmRNA.

H97230 has an 'A' substituted for a 'C' at bp170. H57912 has a 'G' substituted for a 'C' at bp350. H61357 has a 'C' substituted for a 'T' at bp1193. All three substitutions are in high-quality areas of the EST read and are at least 25bp away from the end of the read. The base substitutions can be attributed to one of three causes: (i) A misread of the gel or artifact, (ii) an error in the PCR amplification of the mRNA, or (iii) a SNP in RNA expressed from the genome of the tissue used to create the library. Cause (ii) is unlikely as PCR is a relatively high-fidelity method of amplifying nucleic acid sequences, and was not investigated. To differentiate between (i) and (iii) we attempted to determine if the noted changes mapped to known p53 mutations. Trace file data that supported the fidelity of our conclusions was also incorporated (trace file data not shown).

Using this information, the potential SNPs described above were mapped to codons 12 (H97230, 'CCC'/Pro → 'CAC'/His) and 72 (H57912, 'CCC'/Pro → 'CGC'/Arg).

1) The discrepancy between EST H61357 and wild type at Ala353 (→Val353) is either a base calling error or is a silent mutation not reported in p53/cancer database [7]. Alanine to valine is not usually considered a disruptive mutation.
2) The discrepancy between EST H97230 and wild type at Pro12 (→His12) corresponds to reported mutations [7]. The discrepancy is in an evolutionary conserved region. Finally, H97230 is from a cancerous tissue [8]. Evidence is suggestive that this discrepancy corresponds to a true SNP and not a base calling error.
3) The discrepancy between EST H57912 and w/t at Pro72 (→Arg72) corresponds to mutations reported in both p53 DB and SwissProt [9]. Not only is a mutation reported at this position, but the replacement amino acid is the same as reported in SwissProt (i.e., SwissProt variant 72 is P→R). However, H57912 is not from a disease tissue (Soares: fetal liver/spleen).
4) Although several copies of an EST cSNP are preferable, the data is suggestive in this case.

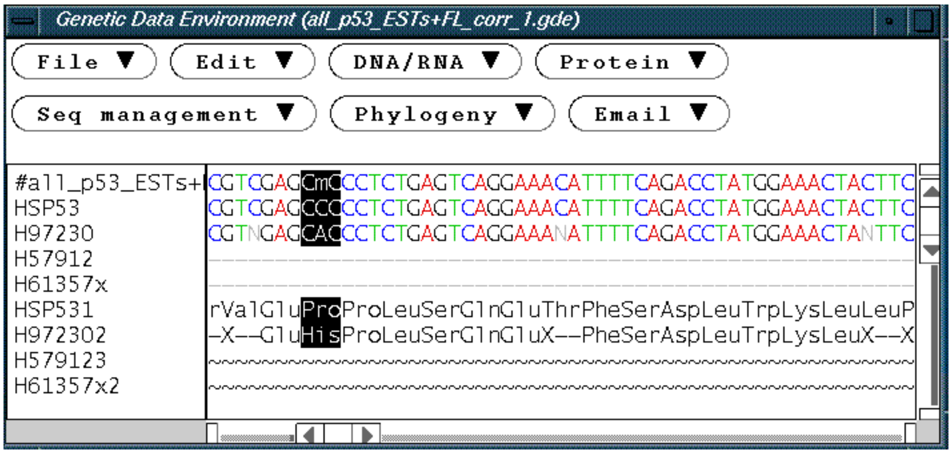

Figure 1

5 Discussion

Our study reveals that mining of EST alignments can produce a number of potential cSNPs. Potential polymorphism(s) in protein P53 were identified and suggestive evidence to support polymorphism was determined. Although it is clear that cSNPs can be detected using the available EST alignments, reliable discrimination methods between false positives, known mutations, and uncharacterised mutations need to be determined. Potential cSNPs can be isolated but the degree of false positives has yet to be identified. Statistical support, empirical data evidence and laboratory confirmation of polymorphisms is therefore necessary. We are in the process of deriving a total number of potential cSNPs, a total number of "known" mutations, and a means to cross-reference the information so that an accurate estimate of total potential cSNPs can be made. The system has shown promising utility when comparisons with known disease causing genes have been employed.

Acknowledgments

We acknowledge the support of the South African Foundation for Research Development.

References

[1] Hide, W., Burke, J., Christoffels, A., Miller, R., A novel approach towards a comprehensive consensus representation of the expressed human genome, *Genome Informatics 1997*, Universal Academy Press, 187–196, 1997.

[2] Gu, Z., Hillier, L., and Kwok, P.Y., Single nucleotide polymorphism hunting in cyberspace, *Hum. Mutat.*, 12(4):221-225, 1998.

[3] Boguski, M.S., Lowe. T.M., and Tolstoshev, C.M., dbEST–database for "expressed sequence tags", *Nat. Genet.*, 4(4):332–333,1993.

[4] http://www.ncbi.nlm.nih.gov/Entrez/

[5] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman DJ., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[6] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215(3):403–410, 1990.

[7] Hainaut, P., Hernandez, T., Robinson, A., Rodriguez-Tome, P., Flores, T., Hollstein, M., Harris, C.C., and Montesano, R., IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools, *Nucleic Acids Res.*, 26(1):205–213, 1998.

[8] Lennon, G., Auffray, C., Polymeropoulos, M., Soares, M.B.., The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression, *Genomics*, 33(1):151-152, 1996.

[9] http://www.ebi.ac.uk/ebi_docs/swissprot_db/relnotes34.html