Genome analysis

Genomic mutation consequence calculator

John E. Major*

Memorial Sloan-Kettering Computational Biology Center, 1275 York Avenue, Box 460, New York, NY 10021, USA

Received on May 15, 2007; revised and accepted on June 19, 2007

Advance Access publication June 28, 2007

Associate Editor: John Quackenbush

ABSTRACT
Summary: The genomic mutation consequence calculator (GMCC) is a tool that will reliably and quickly calculate the consequence of arbitrary genomic mutations. GMCC also reports supporting annotations for the specified genomic region. The particular strength of the GMCC is its ability to perform sequence analysis on regions affected by the mutation. The processed UCSC and COSMIC data resulted in 11 sequences being analyzed with various techniques (http://www.softgenetics.com/ms/index.htm; Chen et al., 2007; Nickerson et al., 1997; Zhang et al., 2005). Studies to date have not always been clear as to how they handle mutation effects in alternate gene isoforms, and some studies have only reviewed mutation effects in a single representative gene isoform. The oversight in using only a single isoform as a reference is that no reports are made on the effects of the mutation in the alternate splice forms. Mutations that appear synonymous in a single reference isoform can often have non-synonymous effects in an alternate isoform (Fig. 1). The variants found in these and future studies are all genomic mutations, and the effects of which should not be simplified to a representative mRNA sequence, but handled in their genomic context.

We built GMCC to report a more complete range of possible effects for a given mutation. Mutations, both point and indel, may be specified anywhere in a genome. GMCC will then report a variety of genomic effects the mutation could cause. GMCC will report if the mutation is upstream of a known gene, whether there is a dbSNP entry for that location, and the conservation score at the specified position. If the mutation hits a gene, GMCC will report the effect upon all gene isoforms including splice site, UTR and coding effects. Additionally, several annotations are gathered for the specified mutation: known COSMIC mutations (Forbes et al., 2005), disease associations, reactome data and Interpro domains provided by the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgTables).

Availability: GMCC is freely available via a web portal with a manual mode and a batch query mode. It may be found at this URL: http://cbio.mskcc.org/gmcc

Contact: majorj@mskcc.org

Supplementary information: A FAQ and examples can be found at the URL above.

1 INTRODUCTION
Genomic resequencing projects with the goal of discovering point mutations and small indels are becoming increasingly common (Bignell et al., 2006; Davies et al., 2005; Greenman et al., 2007; Sjoblom et al., 2006; Stephens et al., 2005; The International HapMap Consortium, 2005). With the NIH’s the cancer genome atlas (TCGA) initiative (http://www.genome.gov/cancersequencing/) underway there will be substantially more resequencing studies in coming years, and a standardized and robust method of describing the effects of genomic mutations is needed. Current resequencing studies fall into two general categories: those that focus on elucidating variation in the genome, and those that focus on trying to discover variation linked to specific diseases. The disease-focused studies share many similarities. They tend to target coding exons of genes using PCR, and the PCR products are sequenced using Sanger sequencing. The resulting sequences are then analyzed with a variety of techniques (http://www.softgenetics.com/ms/index.htm; Chen et al., 2007; Nickerson et al., 1997; Zhang et al., 2005). Studies to date have not always been clear as to how they handle mutation effects in alternate gene isoforms, and some studies have only reviewed mutation effects in a single representative gene isoform. The oversight in using only a single isoform as a reference is that no reports are made on the effects of the mutation in the alternate splice forms. Mutations that appear synonymous in a single reference isoform can often have non-synonymous effects in an alternate isoform (Fig. 1). The variants found in these and future studies are all genomic mutations, and the effects of which should not be simplified to a representative mRNA sequence, but handled in their genomic context.

2 METHODS
GMCC is comprised of three layers. Source data is stored in a Postgres database. Access to this database, and enforcement of all business rules, is managed by a set of Perl modules. Finally, the interface to the application is provided using the Perl CGI::Application package. The web interface offers a manual single query mode as well as a batch query mode.

The source data that supports GMCC is largely drawn from the UCSC Genome Browser ‘Tables’ (http://genome.ucsc.edu/cgi-bin/hgTables). The following UCSC tables were used in building the GMCC database from builds hg17 and hg18: knownGene, snp, knownToLocusLink and multiz17. From proteome: spReactome, spDisease, interProXref. Several tables were denormalized to create GMCC source tables which would be more efficient to query. The COSMIC database (Forbes et al., 2005) was stored as a GMCC source table.

The processed UCSC and COSMIC data resulted in 11 GMCC tables (see Supplementary Material). These tables contain in
Fig. 1. Two distinct non-synonymous coding effects on four isoforms of MAPK9 for the single genomic point mutation of position chr5 179596124 (C->A genomic, G->T coding) in hg18.

total ~55 million rows which required extensive optimization to allow for reasonable query execution times. Queries against this database fall into two categories: queries retrieving data using a unique key, and queries of features on a specific chromosome within set start and end coordinates. The first query type can be easily optimized by utilizing indexes on columns being queried. Defining all features on a chromosome using the built in Postgres geometric data type called ‘box’ optimizes the second class of query. All features have a numeric depth assigned to them based on their chromosome name, and their start and end coordinates were used as the start and end of the box. Building an R-tree index on these ‘box’ fields dramatically speeds up queries for features. Before using the box data type single queries of these ~55 million rows would often take 10–15 s to run, after implementing the box data types single query speeds dropped to 0.5–2 s.

A Perl module was developed to manage the interactions with the database. This module is responsible for formatting the data retrieved from the GMCC database, as well as calculating the effects of mutations. For example, it determines the codons present in all isoforms at the mutation position, and calculates the mutated codons, as well as their amino acid translations.

GMCC offers two web interfaces. The manual single query mode will allow users to enter single mutations and are presented with a mutation summary page. If the mutation affects the mRNA of a gene isoform, a further detail page is available. The batch mode interface returns all of the information available via the single query mode, however, all results are returned in a single comma delimited text file.

3 RESULTS

GMCC is a tool to quickly and reliably determine the effects of any specified genomic mutation, as well as annotate the region of mutation with other useful genomic information. GMCC has two significant strengths derived from this design. First, the tool will report details of mutations not just in coding regions, but also in UTR, splice sites, introns and intergenic space. Second, it is not blind to effects a given mutation might have in alternate gene isoforms. GMCC is particularly well suited for use in large mutation screens. The aim of GMCC is to offer a complete picture of the predictable effects of any given mutation. The output of GMCC can be used to nominate candidate mutations to be explored in more depth either through validation, or through processing of more rigorous functional effect prediction software (e.g. SIFT, Polyphen). GMCC will standardize, automate and increase the efficiency of coding effect prediction and annotation routinely required during mutation analysis. This will result in more consistent results, and a significant savings in time for both small and large scale users.

ACKNOWLEDGEMENTS

The author would like to thank Alex Lash, Thomas Landers, Barry Taylor and Maureen Higgins for their input and support. Conflict of Interest: none declared.

REFERENCES


