HVP Recommended System Submission: Cafe Variome

The System

Extensive amounts of intrinsically useful mutation data cannot be openly shared for legal, ethical or competitive reasons. The full value of these data is therefore never realised, and this fundamental ‘data access’ problem cannot be ‘solved’ by any technically practicable solution.

The Cafe Variome (CV) approach changes the nature of the problem, by converting it to the challenge of enabling fully open and comprehensive ‘data discovery’ (i.e., making the existence rather than the substance of the data openly accessible).
CV is not a database, but instead aims to be a 'shop window' for openly searching/discovering what data exist in the world, including open access, controlled access, and (currently) unavailable content.

The system aims to allow users to openly search the content of the data that the owner has made discoverable, in sophisticated ways, and thereby determine whether a record of interest does or does not exist in an information resource.

Users of the system can subsequently seek to access or browse the hit data (according to pre-set permissions) under one of the following conditions:

1. Open Access: like a journal, where variant records are publicly available for access directly via Café Variome
2. Linked Access: only reports the existence of a variant and the user is linked to the source database to access the full record
3. Restricted Access: the user requests access to variant records from the data owner/submitter. The user must either belong to a pre-approved group or must request access from the data owner to access the variant record(s)

Features

CV offers a complete data sharing software solution based upon enabling the 'open discovery' of data (rather than data 'sharing') for example, between networks of diagnostic laboratories or disease consortia that know/trust each other and share an interest in certain causative genes or diseases.

An intuitive administrator dashboard gives owners complete control over their data and installation. Dashboard configuration options include:

- Content management system for adding/editing custom pages and menus
- Full control over site appearance (logo, colours, backgrounds, themes)
- Template-based import system for source variant data (Excel, tab-delimited and other formats supported) or via direct submission from diagnostic software (currently supported by Gensearch and Alamut)
- Control over which fields are made discoverable to others and how search results are displayed
- Comprehensive access-control system for users and groups
- Fully documented API to allow other systems to programmatically communicate with a CV instance
- Multiple CV instances can be connected together to form a federated network to allow searching across nodes

The discovery interface allows users to form complex queries to interrogate and discover data across the installation. Subsequent access to full datasets is controlled by the sharing policies outlined above. Sharing polices of the variants in the installation can be controlled by the owner per dataset or on a more fine-grained, per variant level, giving complete control over who can access what.

Each variant in the system can be annotated with any number of terms from any of the NCBO BioPortal phenotype ontologies. This flexibility allows a variant to be associated with a single disease term, or a complex combination of Entity and Quality (EQ) phenotype descriptions. An admin tool generates an up-to-date searchable term tree for all ontologies used in the annotations. This functionality makes use of the BioPortal API to ensure that the latest version of all ontologies, and associated terms, are available to the user.

Demo site: http://demo2.cafevariome.org

Administrator login: admin@cafevariome.org

Administrator password: admin3412

The extent to which the system is currently being utilised

1. Direct submission of variants to a CV instance is supported by the following diagnostic variant interpretation software:
   - Gensearch (Phenosystems)
   - Alamut Visual (Interactive Biosoftware)
2. Cafe Variome "Central" uses the CV software and aims to make variants from public databases/sources openly discoverable. The system currently holds over 1.5 million variants. Efforts are on going to keep up-to-date with other variant sources as they are released/updated.
3. The CV software is being piloted by a number of groups through collaborations developed with disease consortia and diagnostic networks.

The HVP Standards and Guidelines the system is compliant with

HGVS nomenclature (RSS001) is supported and an HGVS-compliant description is a required field for variant records. Mutalyzer (RSS002) is used in conjunction with HGVS and has been fully integrated into the system to allow users to validate their HGVS descriptions through the administrator interface (either one at a time or in bulk for whole datasets).

VarioML (RSS004, under review) is used as the core data exchange model in the CV API for submission of variants.

CV accepts the output format from LOVD (RSS003) as an input format. This can be useful for making the core data of an LOVD installation discoverable via a CV instance, with accompanying links back to the full LOVD record to access detailed information.

The LRG standard (RSS005) is supported; users can specify a corresponding LRG for entries. Cafe Variome "Central" has annotated all variants to an LRG if one exists for the entry.

VariO (RSS006, under review) is supported, as it is included in the NCBO BioPortal ontology list. The flexibility of the phenotype system in the CV software means that terms from any ontology included in this list can be used to annotate variants.

The access and license terms and conditions under which the system is made available

The web application software is offered either as a fully hosted solution (on University of Leicester servers) or as "in-a-box" software with a browser-based GUI installer that can be easily deployed on any LAMP type system. Currently, collaborations are arranged on a one-to-one basis. However the software will be made fully available and open-source after publication.