Practical experience with cytogenetics databases and tools

Shuwen Huang
National Genetics Reference Laboratory (Wessex)

06-05-2010 Manchester
Cytogenetics Databases

ACC CVS database
BAC PAC resources
Borgonkar's Chromosome Variation in Man Database
Chromosome Anomaly Collection (CAC)
DECIPHER
Database of Genomic Variants
ECARUCA
ENSEMBL genome browser
Gene & Chromosome Atlas for Oncology & Haematology
Human Telomere Database
Mitelman Database of Chromosome Aberrations in Cancer
NGRI Sub Telomere Collection
OMIM - Online Mendelian Inheritance in Man (NCBI)
ORPHANET - Public database of rare diseases and orphan drugs
UCSC genome browser

Any problems or comments, please email web@cytogenetics.org.uk
International Standard Cytogenomic Array Consortium

**Important Announcement**

We have updated the ISCA website. Please make sure you clear your browser's (e.g. Internet Explorer, Firefox etc.) cache before you start using the new features.

The International Standard Cytogenomic Array (ISCA) Consortium is a rapidly growing group of clinical cytogenetics and molecular genetics laboratories committed to improving quality of patient care related to clinical genetic testing using new molecular cytogenetic technologies including array comparative genomic hybridization (aCGH) and quantitative SNP analysis by microarrays or bead chip technology.

The ISCA database contains whole genome array data from a subset of the ISCA Consortium clinical diagnostic laboratories. Array analysis was carried out on individuals with phenotypes including intellectual disability, autism, and developmental delay.

**Total cases to date: 15,751**

**Pathogenic abnormalities: 2,756 (17.5% of cases)**
Other useful websites

http://cnv.chop.edu

High-resolution mapping of copy number variations in 2,026 healthy individuals

The Copy Number Variation project at the Children's Hospital of Philadelphia (CHOP) represents an effort to identify all frequent copy number variations (CNVs) that exist in the human genome. Our ongoing research utilizes high-resolution genome-wide scanning and highly accurate computational approaches. To date, CNVs have been analyzed in 2,026 healthy children recruited within the Hospital's network. These efforts have yielded the first high-resolution CNV map of the human genome, providing access to a data source that allows for effective assessment of the role of CNVs in human variation and disease susceptibility.

Review and tutorial: the OpenHelix Blog featured the CHOP CNV web site as the tip of the week on November 4, 2009. Make sure to view the blog entry's video tutorial.

The CNV project at CHOP is described in the manuscript High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications, published by Genome Research.

Genomic coordinates are relative to hg17/May 2004/build 35 to match the manuscript. However, it is possible to download the data in hg18/March 2006/build 36.1 coordinates in MySQL dump format. See also the UCSC liftover tool. (To convert a position using liftover, choose the Original Assembly and New Assembly values, set the Data Format value to Position, paste in the position, and click Submit.)

Please report any errors in the web site or data.
Other useful websites

http://www.genetics.med.ed.ac.uk/suspects

**SUSPECTS**

Click here to go straight to the search page or try an example QTL for obesity on 4p

### Introduction

#### What is Suspects?

The aim of Suspects is to efficiently automate the first steps of the candidate gene approach.

In more depth - Suspects is a system for matching Gene Ontology terms, Interpro domains and gene expression data built on top of the PROSPECTR candidate prioritization system. PROSPECTR uses sequence features to rank genes in order of their likelihood of involvement in disease, with Suspects you can drill down further to rank genes involved in specific complex traits and syndromes.

#### How do I use it?

Go to the search page. Enter the markers or coordinates flanking a region of interest. In the next box, enter either the name of the disease phenotype that you are interested in (i.e. hypertension, arthritis, schizophrenia...) or a list of genes that you believe have the same phenotypic effect as the gene that you are looking for.

Elsewhere on the search page you can choose to search the region around a marker or a gene or to search genes that fulfill certain criteria (i.e. all genes with dopamine receptor domains or all genes with a reference to "alcoholism" in their associated literature.)

A Flash based quickstart guide which gives you a brief tour of the interface is available.

#### How does it work?

Suspects operates on the assumption that genes involved in a complex trait will belong to similar pathways and should thus be more likely to share domains, annotation and patterns of expression.
Other useful websites

http://www.genetics.med.ed.ac.uk/prospectr

News

December / January 2005

- Changed the way that training sets are assembled. Consolidated features and added new analyses (mainly more species). Added CO, Interpro and expression data from Ensembl.
- You can now search for matches in functional annotation by using the SUSPECTS server. You may specify a list of genes (by symbol or identified) or a disorder already in the database. Matches are made on the basis of significant semantic similarity. Genes are still ranked using their sequence features but are then given a bonus for sharing Interpro domains, GO terms or similar expression profiles with a given set of “target” genes.
- Retrained the classifier using data from Ensembl Mart v27.

FAQ

What is Prospectr?

It can be shown that genes implicated in disease share certain patterns of sequence based features like longer gene lengths and broader conservation through evolution.

Prospectr (PRIOritization by Sequence & Phylogenetic Extent of Candidate Regions) is an alternating decision tree which has been trained to differentiate between genes likely to be involved in disease and genes unlikely to be involved in disease. By using sequence–based features like gene length, protein length and the percent identity of homologs in other species as input a classification can be obtained for a gene of interest.

The alternating decision trees outputs a classification ("likely to be involved in disease" or "unlikely to be involved in disease"), a score (which is a measure of confidence in the classification) and a breakdown of which factors contributed most to that score.

Given this score we can also roughly estimate how much more or less likely it is that a particular gene is involved in human hereditary disease.
Other useful websites

http://www.ihop-net.org/UniPub/iHOP
Other useful websites

http://www.possum.net.au/about.html

About Us

POSSEUM-web is a computer-based system that helps clinicians to diagnose syndromes in their patients. It contains information on more than 8000 syndromes, including multiple malformation syndromes, chromosomal deletions and duplications, skeletal dysplasias and metabolic conditions with dysmorphic features. The comprehensive medibase with extensive clinical photos for most conditions, also includes x-rays, diagrams, and histopathology slides.

Using POSSEUM-web, clinicians can search for syndromes based on a patient’s traits or by syndrome name to assist them in making a diagnosis or to learn about syndromes.

Syndrome commentaries provide detailed information about clinical attributes, differential diagnoses, radiology and genetics. Our extensive trait dictionary includes a searchable atlas to assist in choosing the most appropriate trait to describe your patient.

We have direct links to OMIM (Online Mendelian Inheritance in Man). You will need an active internet connection to use POSSEUM-web. The database is updated continuously and the data uploaded every month. The images are updated with annual renewal of your subscription.

The POSSEUM team is based at the Victorian Clinical Genetics Service and The Murdoch Children’s Research Institute Melbourne.

The POSSEUM team acknowledges the contribution of Professor Agnes Bankier over many years.

You can contact us at any time on possum@monash.edu.au

- Dr Cathie Rose – Curator of the database
- Mr Stephen Dwy – Business Manager
- Mr Stanley Ho – IT support and development

What is new about POSSEUM Web?

We are continuously adding new information and new images to POSSEUM-web. Since POSSEUM-web became available over 250 syndromes and more than 1000 new images have been added.

We are currently reviewing the needs of our subscribers and we are developing a wider range of available subscriptions.
An ontology is a computational representation of a domain of knowledge based upon a controlled, standardized vocabulary for describing entities and the semantic relationships between them. The Human Phenotype Ontology (HPO) aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease. Terms in the HPO describes a phenotypic abnormality, such as atrial septal defect.

The HPO was initially developed using information from Online Mendelian Inheritance in Man (OMIM), which is a hugely important data resource in the field of human genetics and beyond. The HPO is currently being developed using information from OMIM and the medical literature and contains approximately 10,000 terms. Over 50,000 annotations to hereditary diseases are available for download or can be browsed using the PhenExplorer.

The HPO is now being developed in collaboration with members of the OBO Foundry (Open Biological and Biomedical Ontologies), and logical definitions for HPO terms are being developed using PATO and a number of other ontologies including the FMA, GO, ChEBI, and MPATH. The HPO can be used for clinical diagnostics in human genetics (Phenomizer), bioinformatics research on the relationships between human phenotypic abnormalities and cellular and biochemical networks, for mapping between human and model organism phenotypes, and for providing a standardized vocabulary for clinical databases, among many other things.

The HPO project encourages input from the medical and genetics community with regards to the ontology itself and to clinical annotations.

Other useful websites

HPO (Human Phenotype Ontology)

Other useful websites


Endeavour

Candidate genes prioritization through genomic data fusion

You are here on the web client version of Endeavour. For an introduction, latest news, academic or commercial use, contact info, mailing list, please refer to the Endeavour project main page.

How to cite

  Abstract

  Abstract

Prioritize candidate genes

A manual is available here. If this is your first visit, you may try out some examples:

- YPEL1 taken from our Nature biotech. paper on DiGeorge syndrome candidate genes.
- KCNJ5 taken from the Elbers et al review on obesity and diabetes.
- DFNB31 based on the Fbermann et al paper describing the discovery of a novel Usher gene.

Clicking on the gene name (YPEL1, KCNJ5 or DFNB31) will cause the training and the candidate genes to be loaded in the following wizard; this will also select the most appropriate data sources. Then you should go over the different steps (Next' button, enabled once the training and candidate genes are loaded) to revise the settings and launch the prioritization.

Prioritize your candidates in 4 steps with the following wizard
Other useful websites

http://biogps.gnf.org/#goto=welcome
Cytogenetics tools (for microarray)

• Design your own custom array — eArray
  https://earray.chem.agilent.com/earray/

• Array analysis and interpretation
  — Genomic Workbench (Agilent)
  — BlueFuse Multi (BlueGnome)
  — CytoSure (OGT)
  — Genoglyphix (Signature Genomics)
  — ChAS (Affymetrix)
  — Cartagenia BENCH (Cartagenia)
  — ......
Design your own custom array
Array design - why

Example 1: some regions are not covered by catalog array.

PTCH1 gene is related to Gorlin syndrome
Example 2: some regions might be missed due to fewer probes

A2BP1 is a reported candidate gene for autism
Example 3: the coverage of some regions are not convincing enough

TCF4 is related to mental retardation.
Solution – design your own array

-Agilent eArray software

-https://earray.chem.agilent.com/earray/
- Free registration
- Create custom array designs
- High density database
- Collaboration and design sharing
- Access all Agilent catalog content
- Visualisation of probe distribution
- Useful tools
eArray provides an easy way of managing the array creation process.
BED  $\rightarrow$ UCSC, Ensembl

Design file $\rightarrow$ Analysis software packages
Checking or visualising probe distribution

**UCSC Genome Browser**

http://genome.ucsc.edu/cgi-bin/hgGateway

**Ensembl**

http://www.ensembl.org/Homo_sapiens/Location
UCSC

- Genome browser
- Manage custom track
- Add custom tracks
- Upload BED file and submit
- Go to genome browser
- Check probe distribution by chromosome location or by genes
UCSC basic settings

- Base position
- Chromosome band
- FISH clones
- UCSC genes
- DGV Struct Var
- Segmental Dups
- Structural Var

Click each section for more detailed information
Analysis and interpretation
Genomic Workbench

• Agilent Technologies
• Updated version: Genomic Workbench 6.0
• Multi-functional:

- eArray_{XD}: available
- License is needed, free trial is available at:
ChAS

- **Chromosome Analysis Suite (ChAS)**
- Affymetrix
- For 2.7M Cytogenomics Solution array
- Free downloadable
- Detection of copy number variants, and **Long Contiguous Stretches of Homozygosity (LCSH)**
Genoglyphix

- Software from Signature Genomics
- In collaboration with NimbleGen
- Web-based software
- [https://www.genoglyphix.com](https://www.genoglyphix.com)
- License is needed
- Only suitable for Signature Genomics’ design of NimbleGen array
Bluefuse Multi

- BlueGnome
- Manage array information
- Automated data analysis
- Decision Track visualization
- Unlimited license free of charge when purchasing CytoChip microarrays
- Brings together information from multiple data sources to support follow up decision.
  - Underlying information provided and preserved in annotation DB.
CytoSure

- Powerful data analysis and visualisation tool
- Determines if the aberration is causative:
  - Automated aberration detection
  - Eliminate non-pathogenic CNVs
  - Track quality metrics
  - Link to publicly available data
  - Get functional information on the genes
  - Determine if aberration is *de novo*
  - Compare to previous data
- All aberrations and annotation saved to centralised sample database
- Legacy data visible as an extra track
.txt file

Import .txt file using File -> Import

Initial view of data

Run CBS to identify aberrations

Manually verify aberrations

Classify the aberrations identified

File -> Save ➔ .cgh file

Save aberrations and annotation to database ➔ database

Export results for printing
Cartagenia BENCH

- Data interpretation and data sharing
- Automatic upload to DECIPHER
- Detects and draws attention to similar patients
- Intelligent search, clustering, and diagnosis support
  - Find e.g. patients and CNV’s linked to any heart phenotype
- Gene prioritisation // ENDEAVOUR
- Automatically identify which patient CNV’s are linked to patient phenotype
- Suggests literature that is relevant in patient context
- License is needed
Manage CNV sets: ‘your own DGV’

- Build & use CNV sets from patients / controls
  - Private
  - Shared
  - Custom
  - Public
    - DGV
    - CHOP
    - ...

[Image of a graphical user interface showing the management of CNV sets]
Intelligent CNV filtration

Filter regions associated with case report(s)

- ChromosomeFilter
- CNVFilter
- GeneContentFilter
- KnownSyndromeFilter

Number of regions in input: 8 from 1 different samples

- Advanced region input and merging:
  - Search case reports
  - Search file
  - Merge regions

- Filter regions by:
  - Length
  - Chromosome
  - Overlap with known CNVs
  - Base content
  - Based on known syndromes
  - Inheritance
  - In other regions
  - Against other regions
  - Repetitive
  - Against other regions
  - Using custom filters

Remaining regions
- Go to manual review
- Check excluded regions

Excluded regions
- Bulk edit remaining regions
- Bulk edit excluded regions

Label remaining regions
- Label excluded regions

Export remaining regions to Excel
- Export excluded regions to Excel
Integration with data analysis tools

**Cartagenia BENCH track**

Visualize info from several BENCH installations in OGT CytoSure

One-click export to BENCH platform
Hands on time