Database Integration

Paul Flicek
Vertebrate Genomics
(Dramatically) Simplified Clinical Workflow

Identify variants

Technically easy and getting easier

Use what we already know to make some sense of them

Do something about it
Data interpretation: beyond research toward medical practice

• Needs:
  • Consistent, traceable data generation and analysis routines
  • Robust annotation based on public information sources such as those at the EBI and NCBI
    • Probably 95% of all information that could be used to understand and interpret human variation is already in the public domain
  • Reporting into medical records
Database integration

- Part 1: Continually update the existing information to ensure it is accurate and comprehensive

- Part 2: Provide some method to search relevant resources using variants and/or whole genomes as input
The European Genome-phenome Archive (EGA)

The European Genome-phenome Archive (EGA) repository allows you to explore datasets from numerous genotype experiments, supplied by a range of data providers.

### Studies

Studies are experimental investigations of a particular phenomenon or trait.

- [Browse all studies](#)

### Datasets

The EGA archives a large number of datasets, the access to which is controlled by a dataset access committee (DAC).

- [Browse datasets that we hold](#)

### Data Providers

Data Providers can be involved in creating studies, data submission and the designation of data access committees (DACs).

- [Browse EGA Data Providers](#)

### Learn about the EGA

- [Introduction to the EGA](#)
- [How to obtain an account with the EGA](#)

### Video resources

- [Watch a video of how to use your EGA account](#)

### Help

- [Users FAQ](#)
- [Submitters FAQ](#)
- [FTP & Aspera FAQ](#)
- [Contact Us](#)
- [EGA mailing list](#)
The European Genome-phenome Archive

- Secure storage and authorised access to all types of data sets that might be generated in the context of research into molecular medicine
  - DNA sequence; Array-based genotypes; epigenetic data
  - Transcriptomics; Proteomics
  - Phenotype data

- Used for GWAS, ICGC, IHEC, IHMC, UK10K and data

- EGA supports only data access decisions that are based on original consent
  - Authorized users have personal accounts in our system
  - Access to the data requires account password
  - Data decryption requires a separate key that must be requested and is sent off line
Ensembl genome-wide annotation

- Genomic alignments
- Chromosomes
- Genes
- Synteny
- Gene families
- SNPs
- Across species
- Within species
- Orthology
Integrating variation data across the genome

- Polymorphism data (from dbSNP)
  - SNPs and indels for 14 species including 1000 Genomes
  - Allele and genotype frequencies by population
  - Locus-specific data from LRG
- Structural polymorphism data
- Mutation data (human)
  - Somatic mutation data (from COSMIC)
  - Human Gene Mutation Database (HGMD) IDs
- Phenotype associations: OMIM, UniProt, GWAS
- Affymetrix and Illumina chipsets
Variation annotation – phenotype data

- 37,964 somatic mutations:
  - COSMIC
- 57,930 germline mutations:
  - HGMD
- 56,177 literature curation:
  - OMIM
  - UniProt
- 62,737 GWAS data:
  - NHGRI GWAS catalog
  - Open Access DB
  - EGA
- 22,449 from SNPedia by DAS
Variation annotation – phenotype data

- LRG project - Locus Reference Genomic
  - Create stable reference sequences (LRGs)
  - Use LRGs for exchange of variation data

LSDBs
Diagnostic labs
Locus-specific information

Genome-wide information

Dalgleish, et al. Genome Medicine 2010
Database integration

• Part 1: Continually update the existing information to ensure it is accurate and comprehensive

• Part 2: Provide some method to search relevant resources using variants and/or whole genomes as input
Ensembl Variant Effect Prediction (VEP) tool

- Calculates the effect of SNPs in the context of Ensembl genes and regulatory features
  - Web and API interface
  - Code back-ported to support NCBI36 assembly
  - Programmatic support for tab-delimited and VCF files
  - Easily integrated into analysis pipelines
- Working within ICGC to capture structural and other genome rearrangements
- Disruption of experimentally observed TF binding sites and conserved regions
- Ability to run without connection to the internet
- Support for user defined analysis plug-ins coming in January 2012
- Will return if variant is present in EGA dataset in 2012
- Effectively a variant based search of EBI’s data resources

Ensembl VEP Implementation

```perl
#!/usr/bin/perl
use strict;
use Bio::EnsEMBL::Registry;
use Bio::EnsEMBL::Variation::DBSOL::VariationFeatureAdaptor;

# get registry
my $reg = Bio::EnsEMBL::Registry;
$reg->load_registry_from_db;
-host => "ensembldb.ensembl.org",
-user => "anonymous"
);
my $sa = $reg->get_adaptor('human', 'core', 'slice');
my $sa = $reg->get_adaptor('human', 'variation', 'variationfeature');
my $slice = $sa->fetch_by_region('chromosome', 13);
my $vf = Bio::EnsEMBL::Variation::VariationFeature->new;
-variation_name => 'Var1',
-slice => $slice,
-start => 1294818,
-end => 1294818,
-strand => 1,
-allele_string => 'AVG',
-adaptor => $vf;
my $svs = $vf->get_all_TranscriptVariations();
```
50+ species at [www.ensembl.org](http://www.ensembl.org)
300+ at [www.ensemblgenomes.org](http://www.ensemblgenomes.org)

Data input by file upload or external URL

Support for multiple file formats: VCF, Pileup, HGVS, dbSNP rsID

Output Ensembl, Sequence Ontology (SO) or NCBI consequence terms

Find existing overlapping variants annotated by Ensembl

Create HGVS notations

Include SIFT, PolyPhen and Condel predictions for non-synonymous changes in human

Filter input against HapMap or 1000 genomes frequency data
<table>
<thead>
<tr>
<th>Uploaded Variation</th>
<th>Location</th>
<th>Allele</th>
<th>Gene</th>
<th>Feature</th>
<th>Feature type</th>
<th>Consequence</th>
<th>Position in cDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>C</td>
<td>300000154719</td>
<td>ENST0000307301</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>915</td>
</tr>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>C</td>
<td>ENSG0000154719</td>
<td>ENST0000419219</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>852</td>
</tr>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>C</td>
<td>ENSG0000154719</td>
<td>ENST0000352957</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>915</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>C</td>
<td>ENSG0000154719</td>
<td>ENST0000307301</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>882</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>C</td>
<td>ENSG0000154719</td>
<td>ENST0000352957</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>882</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>C</td>
<td>ENSG0000154719</td>
<td>ENST0000419219</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>819</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>-</td>
<td>ENSR00000613843</td>
<td>RegulatoryFeature</td>
<td>regulatory_region_variant</td>
<td>-</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>ENSG00000160183</td>
<td>ENST0000553129</td>
<td>Transcript</td>
<td>NMD_transcript_variant, intron_variant</td>
<td>-</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>ENSG00000157557</td>
<td>ENST0000360214</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>1393</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>ENSG00000157557</td>
<td>ENST0000360938</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>1223</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>ENSG00000157557</td>
<td>ENST0000456966</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>1223</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191548</td>
<td>T</td>
<td>ENSG00000157557</td>
<td>ENST0000432278</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>1223</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965148</td>
<td>A</td>
<td>ENSG0000154719</td>
<td>ENST0000307301</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>939</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965148</td>
<td>A</td>
<td>ENSG00000154719</td>
<td>ENST0000419219</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>876</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965148</td>
<td>A</td>
<td>ENSG00000154719</td>
<td>ENST0000352957</td>
<td>Transcript</td>
<td>synonymous_codon</td>
<td>939</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029195</td>
<td>G</td>
<td>ENSG00000159082</td>
<td>ENST0000357345</td>
<td>Transcript</td>
<td>non_synonymous_codon</td>
<td>2722</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029195</td>
<td>G</td>
<td>ENSG00000159082</td>
<td>ENST0000382499</td>
<td>Transcript</td>
<td>non_synonymous_codon</td>
<td>2714</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029195</td>
<td>G</td>
<td>ENSG00000159082</td>
<td>ENST0000322229</td>
<td>Transcript</td>
<td>non_synonymous_codon</td>
<td>2597</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029195</td>
<td>G</td>
<td>ENSG00000159082</td>
<td>ENST0000464778</td>
<td>Transcript</td>
<td>nc_transcript_variant</td>
<td>384</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029195</td>
<td>G</td>
<td>ENSG00000159082</td>
<td>ENST0000439331</td>
<td>Transcript</td>
<td>non_synonymous_codon</td>
<td>2722</td>
</tr>
<tr>
<td>Uploaded Variation</td>
<td>Location</td>
<td>Reference</td>
<td>Position in cDNA</td>
<td>Position in CDS</td>
<td>Position in protein</td>
<td>Amino acid change</td>
<td>Codon change</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------------</td>
<td>----------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>s.codon</td>
<td>915</td>
<td>873</td>
<td>291</td>
<td>P</td>
<td>ccA/ccG</td>
</tr>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>s.codon</td>
<td>852</td>
<td>843</td>
<td>281</td>
<td>P</td>
<td>ccA/ccG</td>
</tr>
<tr>
<td>rs10576</td>
<td>21:26965172</td>
<td>s.codon</td>
<td>915</td>
<td>873</td>
<td>291</td>
<td>P</td>
<td>ccA/ccG</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>s.codon</td>
<td>882</td>
<td>840</td>
<td>280</td>
<td>V</td>
<td>gtA/gtG</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>s.codon</td>
<td>882</td>
<td>840</td>
<td>280</td>
<td>V</td>
<td>gtA/gtG</td>
</tr>
<tr>
<td>rs1057885</td>
<td>21:26965205</td>
<td>s.codon</td>
<td>819</td>
<td>810</td>
<td>270</td>
<td>V</td>
<td>gtA/gtG</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>1393</td>
<td>933</td>
<td>311</td>
<td>F</td>
<td>ttC/ttT</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>1223</td>
<td>933</td>
<td>311</td>
<td>F</td>
<td>ttC/ttT</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rs113417859</td>
<td>21:40191540</td>
<td>s.codon</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965144</td>
<td>s.codon</td>
<td>939</td>
<td>897</td>
<td>299</td>
<td>G</td>
<td>ggC/ggT</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965144</td>
<td>s.codon</td>
<td>876</td>
<td>867</td>
<td>289</td>
<td>G</td>
<td>ggC/ggT</td>
</tr>
<tr>
<td>rs1135638</td>
<td>21:26965144</td>
<td>s.codon</td>
<td>939</td>
<td>897</td>
<td>299</td>
<td>G</td>
<td>ggC/ggT</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029192</td>
<td>s.codon</td>
<td>2722</td>
<td>2597</td>
<td>866</td>
<td>I/T</td>
<td>aT/aCt</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029192</td>
<td>s.codon</td>
<td>2714</td>
<td>2714</td>
<td>905</td>
<td>I/T</td>
<td>aT/aCt</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029192</td>
<td>s.codon</td>
<td>2597</td>
<td>2597</td>
<td>866</td>
<td>I/T</td>
<td>aT/aCt</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029192</td>
<td>s.codon</td>
<td>384</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>rs114053718</td>
<td>21:34029192</td>
<td>s.codon</td>
<td>2722</td>
<td>2714</td>
<td>905</td>
<td>I/T</td>
<td>aT/aCt</td>
</tr>
</tbody>
</table>
Sequence Ontology consequences

- Provides a structured controlled vocabulary for the description of mutations at both the sequence and more gross level in the context of genomic databases.
SIFT, PolyPhen and Condel in practice

- Store every possible score for every* protein

| A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
| 0.001 | 0.047 | 0.007 | 0.007 | 0.002 | 0.047 | 0.001 | 0.002 | 0.001 | - | 0.007 | 0.007 | 0.007 | 0.002 | 0.002 | 0.001 | 0.094 | 0.017 |
| 0.081 | 0.547 | 0.547 | 0.348 | 0.201 | 0.348 | 0.817 | 0.081 | 0.348 | - | 0.348 | 0.547 | 0.547 | 0.547 | 0.547 | 0.201 | 0.201 | 0.081 | 0.817 | 0.547 |
| 0.007 | 0.191 | 0.007 | 0.002 | 0.094 | 0.017 | 0.094 | 0.047 | 0.002 | 0.017 | 0.094 | 0.017 | 0.017 | - | 0.007 | 0.007 | 0.017 | 0.017 | 0.191 | 0.047 |
| 0.017 | 0.362 | 0.201 | 0.106 | 0.106 | 0.106 | 0.362 | 0.017 | 0.106 | 0.017 | 0.201 | 0.362 | 0.201 | 0.362 | 0.362 | 0.106 | 0.04 | - | 0.677 | 0.201 |
| 0.017 | 0.362 | 0.201 | 0.106 | 0.106 | 0.106 | 0.362 | 0.017 | 0.106 | 0.017 | 0.201 | 0.362 | 0.201 | 0.362 | 0.362 | 0.106 | 0.04 | - | 0.677 | 0.201 |
| 0.007 | 0.191 | 0.007 | 0.002 | 0.094 | 0.017 | 0.094 | 0.047 | 0.002 | 0.017 | 0.094 | 0.017 | 0.017 | - | 0.007 | 0.007 | 0.017 | 0.017 | 0.191 | 0.047 |
| 0.081 | 0.817 | 0.035 | - | 0.547 | 0.081 | 0.547 | 0.547 | 0.081 | 0.201 | 0.547 | 0.201 | 0.201 | 0.081 | 0.201 | 0.081 | 0.201 | 0.817 | 0.547 |
| 0.663 | 0.99 | 0.964 | 0.964 | - | 0.99 | 0.964 | 0.964 | 0.964 | 0.99 | 0.922 | 0.964 | 0.964 | 0.964 | 0.848 | 0.964 | 0.964 | 0.99 | 0.99 |
| 0.081 | 0.817 | 0.081 | 0.081 | 0.547 | 0.081 | 0.348 | 0.547 | 0.081 | 0.201 | 0.547 | - | 0.348 | 0.201 | 0.201 | 0.081 | 0.081 | 0.201 | 0.817 | 0.547 |

- Condel scores are an algorithmic function of SIFT and Polyphen scores
Regulatory region consequences

- Variant within a regulatory feature = RegulatoryFeature
- Variant within a transcription factor binding motif = MotifFeature
- Variant in an “informative position” = HIGH_INF_POS
Has this variant ever been seen before?

• Quickly becoming the most common question in human genomics
  • Incredibly hard to answer

• Nature said (in the October 2010 1000 Genomes issue) that about 2700 genomes had been sequenced and estimate 30,000 by the end of 2011
  • Beyond the those currently in the 1000 Genomes project (~2000) relatively few of these genomes are easily accessible

• There are many more exomes
  • Access here can be a problem as well

• Some data is available under controlled access and the fraction of data in this category is expected to increase
Future

• Ensembl is not a clinical decision support tool and only a fraction of the important resources were presented

• It does show the way forward
  • Comprehensive
  • Versioned
  • Standardized
  • Using controlled terminology
  • Regularly updated
  • Evidence based and algorithmic
  • Fully open

• There is uncertainty at every step in the process from the genome reference to the gene set to the interpretation and we have to work in this environment
Acknowledgements

• Ensembl Annotation and VEP: Will McLaren, Graham Ritchie, Pontus Larsson, Daniel Sobral, Bethan Yates, Anne Parker, Jackie MacArthur, Fiona Cunningham

• EBI Variation Archives: Ilkka Lappalainen, Vasudev Kumanduri, Dylan Spalding, Mick Maguire, Lisa Skipper, Jeff Almeida-King

• Funding: Wellcome Trust, European Commission, NHGRI, British Heart Foundation, EMBL
EBI data integration and added value

- EBI search provides integration into EBI existing spines (DAS based)
- Development of new spines diseases, cell type, tissue, tools
- User focussed design with general and specific user groups
- Added value - terminology, literature searching, pathways etc (user defined)
- Reciprocal integration between KOMP2 web portal and EBI resources
Alkaline phosphatase, liver/bone/kidney Gene

Alpl differential expression summary

Organism part
liver, kidney, placenta, thymus
View all

Disease state
normal, (empty), Pb-A infected, myocardial infarction
View all

Cell type
embryonic stem cell, hematopoietic stem cell, T effector cell, T regulatory cell
View all

Cell line
N1E-115 wild_type, 67NR, Swiss5, Swiss8
View all

Compound treatment
control, none, vehicle - arachis oil, 17beta-estradiol
View all

Developmental stage
embryo, adult, fetus, neonate
View all