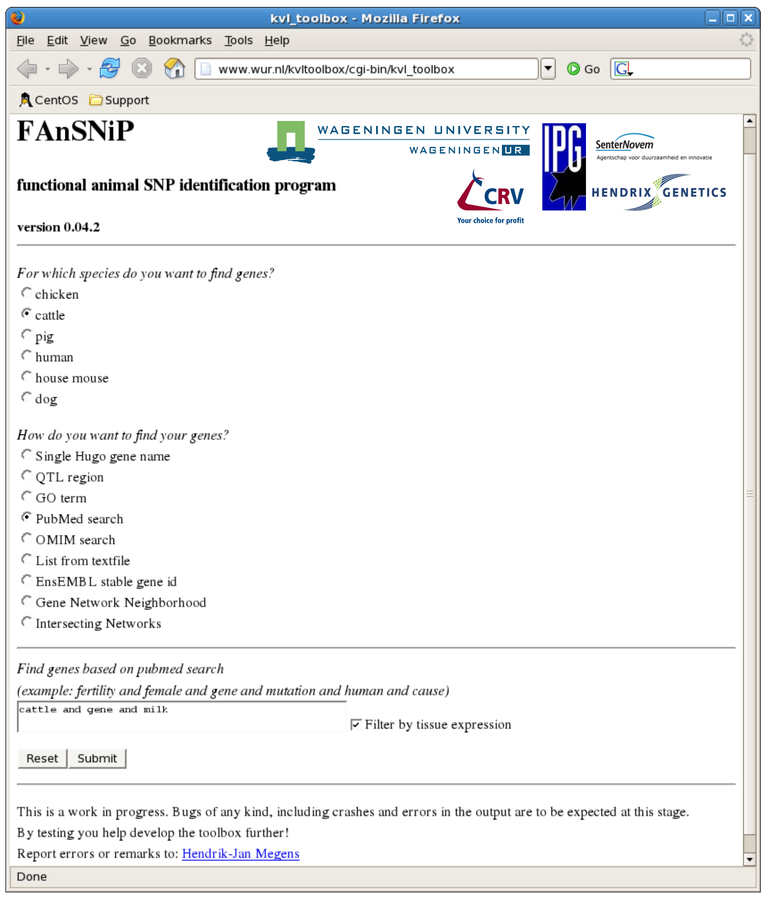
A Perl based data mining workflow for animal breeding - from phenotype to SNP

Hendrik-Jan Megens[1], Richard P.M.A Crooijmans[1], Gerard A.A. Albers[2], Barbara Harlizius[3], Erik Mullaart[4], Sonja Kollers[3], Annemiek P. Jungerius[2], Hindrik H.D. Kerstens[1] & Martien A.M. Groenen[1]

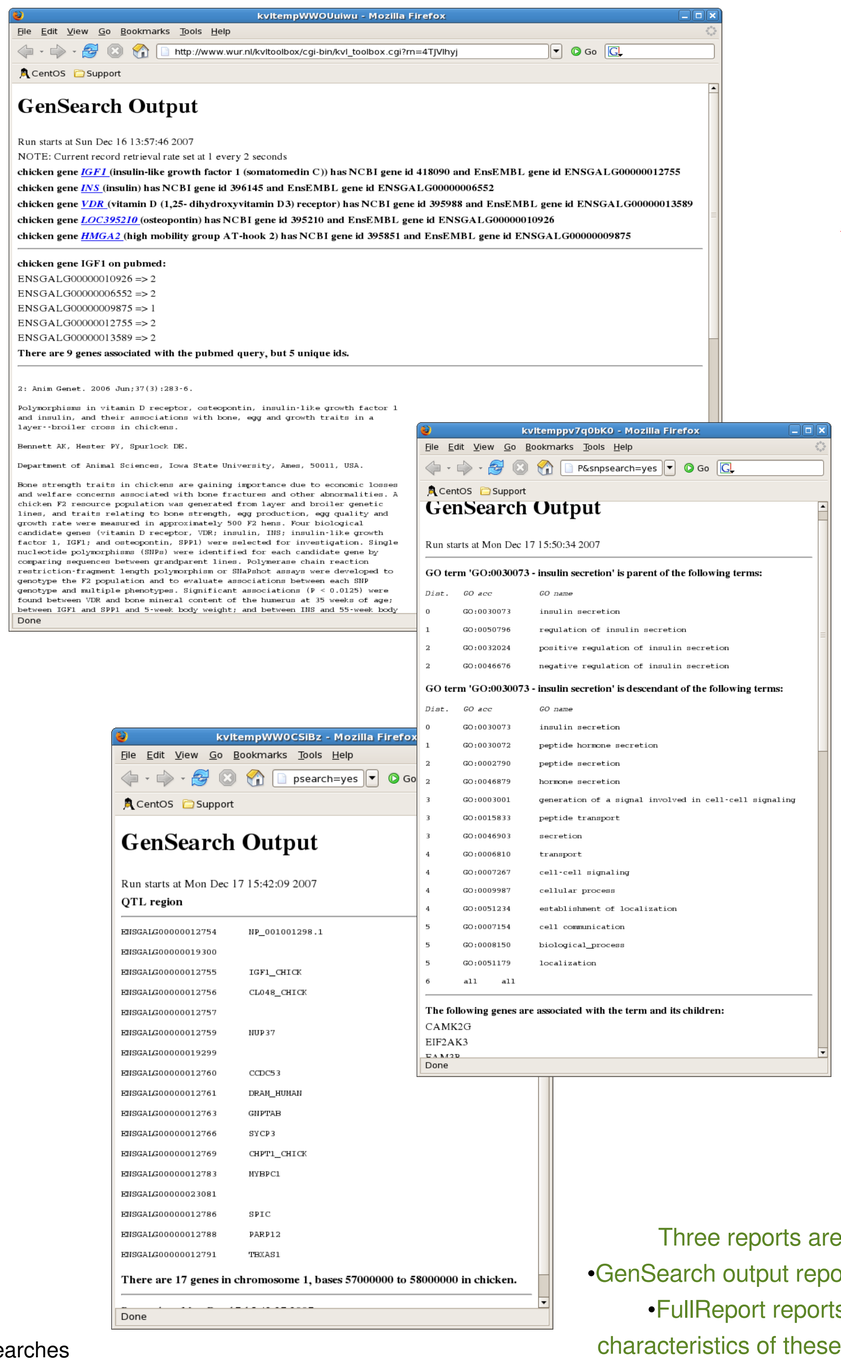
[1] Animal Breeding and Genetics Group, Wageningen University, 6709 PG Wageningen, The Netherlands. [2] Hendrix Genetics P.O. Box 114, 5830AC Boxmeer, The Netherlands.
[3] Institute for Pig Genetics, P.O. Box 43, 6640 AA Beuningen, The Netherlands. [4] CRV, P.O. Box 5073, 6802 EB, Arnhem, The Netherlands
Email: hendrik-jan.megens@wur.nl

With vast and rapidly increasing resources of genomic data available, there is a challenge for breeders to utilize this information in their breeding practices. While tools to combine multiple databases to aid in selecting genes and SNPs for medical studies have become publicly available, none exist that can comprehensively and species specifically do this for livestock animals. We have developed a tool for mining genomic data specifically for breeding purposes. The main objective of the tool is the development of large SNP sets consisting of putative functional SNPs, thus facilitating a 'high throughput' implementation of the candidate gene approach.

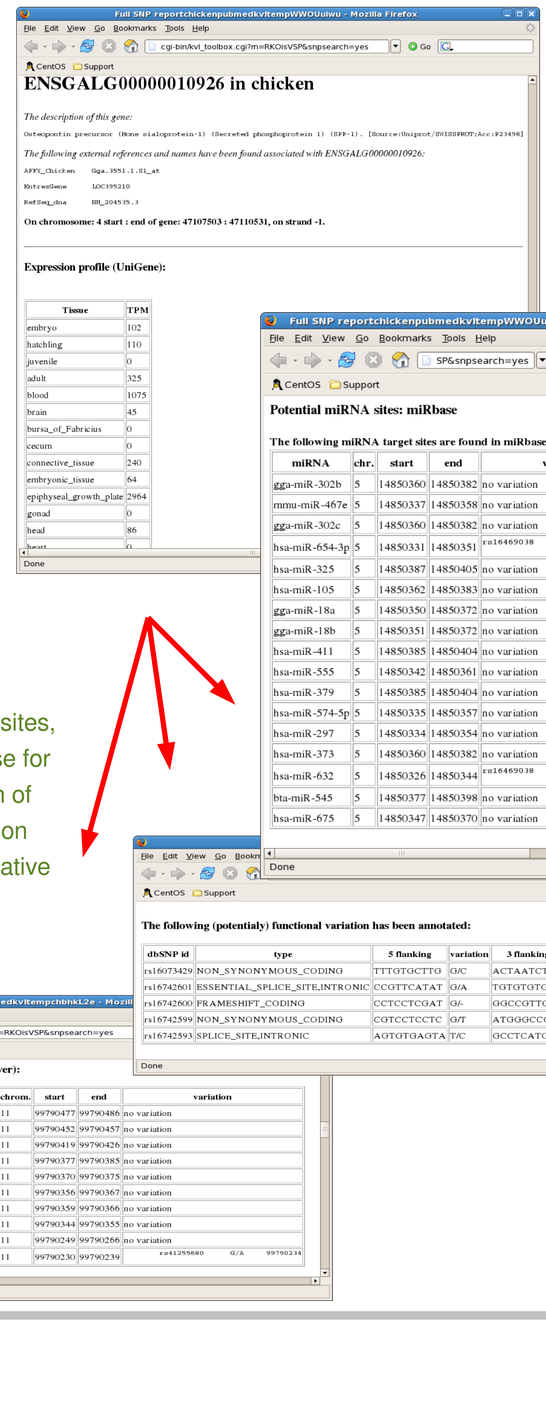
Characteristics of the tool are:
- Perl based pipeline integrating data from livestock and model species
- Modular design which facilitates its use as tool development platform for our genomics research
- Easily extendable to other species with information in EnsEMBL
- Platform independent, CGI based interface providing user-friendly access
- Job submission allows for extensive searches on hundreds or thousands of genes
- Candidate gene selection based on PubMed, GO, gene network information and QTL data
- Identification of putative functional information in candidate genes (TFBS, miRNA)
- Search for known variation (SNPs) at these functional sites



GenSearch based on various query options will result in retrieving a number of genes for one of our favorite organisms



Information on the genes is automatically retrieved, such as various db ids, ortholog info (where applicable), position in the genome, transcript info, expression profiles (UniGene, HPRD) and automatic links to NCBI



Search for annotated functional variation
Search for putative regulatory element binding sites, through information in databases (e.g. MiRBase for miRNA binding sites), and by implementation of various programs (e.g. Clover for Transcription Factor Binding Sites). Retrieve SNP info in putative functional sites through dbSNP

Three reports are generated:
- GenSearch output reports on genes found
- FullReport reports on all useful characteristics of these genes (homology features, positional information, transcripts etc), including the SNP info
- SNPreport reports back just the SNP info

Programming
- Perl 5.8
- Perl CGI, DBI
- BioPerl
- EnsEMBL core Perl API
- EnsEMBL compara API
- EnsEMBL variation API
- NCBI URL API
- MySQL 5.0

Database Resources
- EnsEMBL
- NCBI gene
- UniGene
- miRBase
- Gene Ontology (GO)
- HPRD
- BIND
- BioGRID
- HGNC
- dbSNP
- PubMed
- OMIM

Supporting programs:
- miRanda
- Clover
- FootPrinter
- Blast

Ongoing work:
- Refine species-specific gene searches
- Add expression data from ongoing experiments
- Add functionality for detecting functional regions in genes
- Add company specific information on phenotypes and line specific SNPs

Note: the tool is currently not available without login.