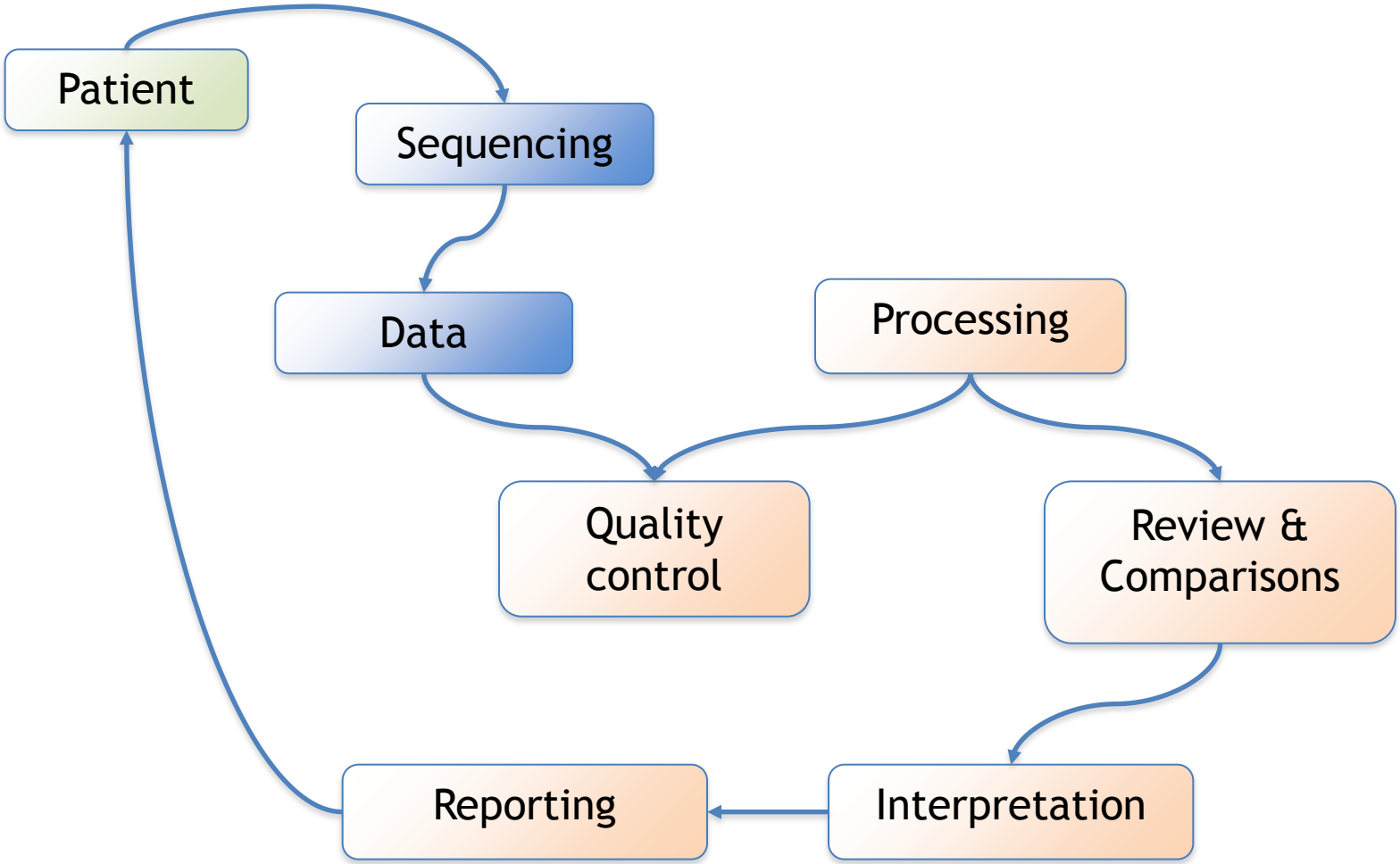
Data Structure & Models

Prof. Andre Franke | Institute of Clinical Molecular Biology | Christian-Albrechts-University Kiel (Germany)

Berlin, 06.12. 2012

ikmb
Institut für Klinische Molekularbiologie

DX vs. Research Workflow



Good Standards Exist

BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

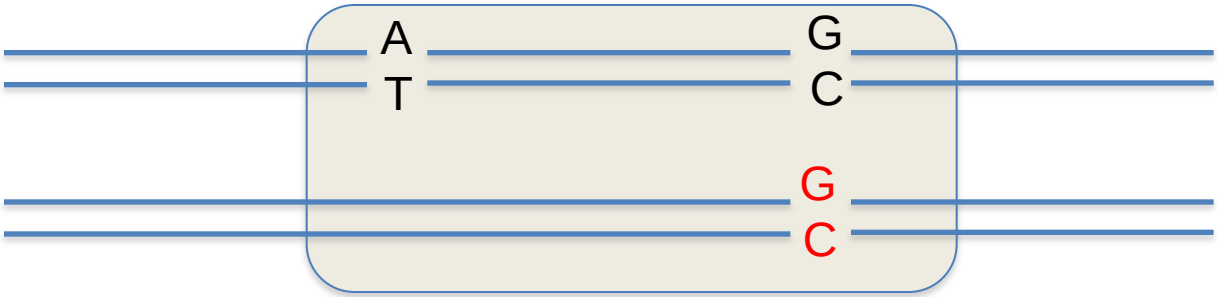
Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

Variant annotation is easy. No!

- Different splice variants
- RefGene vs. ENSEMBL vs. CCDS vs. RefProt
- dbSNP is not tidy
- *in silico* prediction of novel (!) variants
- Most prediction tools are for nonsynonymous SNVs
- Noncoding variants? CNVs? InDels?
- Real vs. not so real compound heterozygotes



piBase – More than a QC tool for SNPs

icmb Institute of Clinical Molecular Biology

pibase

Download | Example Data (12GB) | Example Output (130kb) | QuickStart Tutorial | Contact

pibase tools for validational and comparative analysis of BAM files

pibase is an open-source package of linux command line tools for validating next-generation sequencing loci (SNPs and loci of interest where no SNPs are known) and for comparative analyses using Fisher's exact test.

Acknowledgement: The development of pibase was partly funded by: The German Ministry of Education and Research (BMBF); the National Genome Research Network (NGFN); the Deutsche Forschungsgemeinschaft (DFG) Cluster of Excellence 'Inflammation at Interfaces'; the EU Seventh Framework Programme [FP7/2007-2013, grant numbers 201418, READNA and 262055, ESGI].

Disclaimer: pibase is provided free of charge for non-commercial use but you are required to read our disclaimer and to cite us when publishing results.

Download: pibase 1.4.5 example data (12GB) example output only (130kb)

Overview

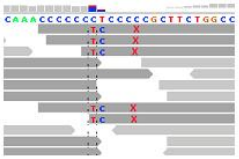
pibase	Acronym for: get **P**osition **I**nformation at **BASE** position of interest.
Interoperability	Input and output file types.
Work flows	Preparing BAM-files and using the pibase tools.
QuickStart Tutorial	Prerequisites, installation, and pibase examples using BAM-files from the 1000 Genomes project.

Essentials:

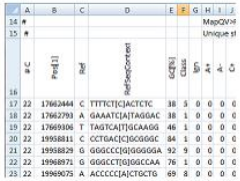
pibase_bamref	Extract information from a BAM-file and a reference sequence file and table this information into a tab-separated text file.
pibase_consensus	Infer 'best' genotypes and their 'quality' classification, and optionally merge multiple pibase_bamref files (e.g. a control panel or several runs of the same patient) into a single file.
pibase_fisherdiff	Compare two pibase_consensus files using Fisher's exact test on original data (aligned reads) rather than comparing processed data (SNP-calls or genotypes).

Annotate:

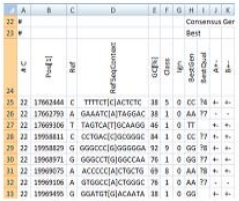
Step 0: Filtering examples



Step 1: BAM / Ref



Step 2: Genotype / Quality

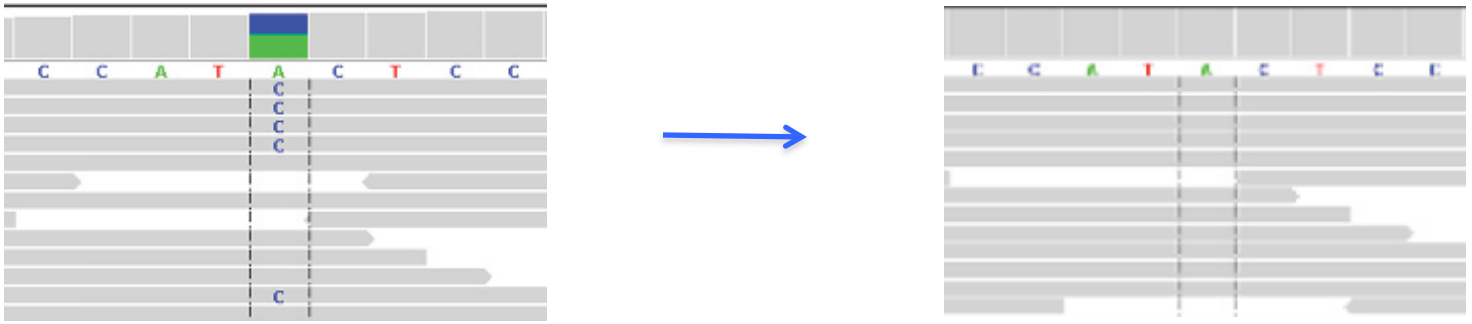


Step 3: Comparison

http://www.ikmb.uni-kiel.de/pibase/ Forster *et al.*, Nucleic Acids Res. 2012

Backmapping

- Target mapping, then backmapping → reduces false-positives while maintaining low false-negative rate



Research article

Highly accessed | Open Access

Improving mapping and SNP-calling performance in multiplexed targeted next-generation sequencing

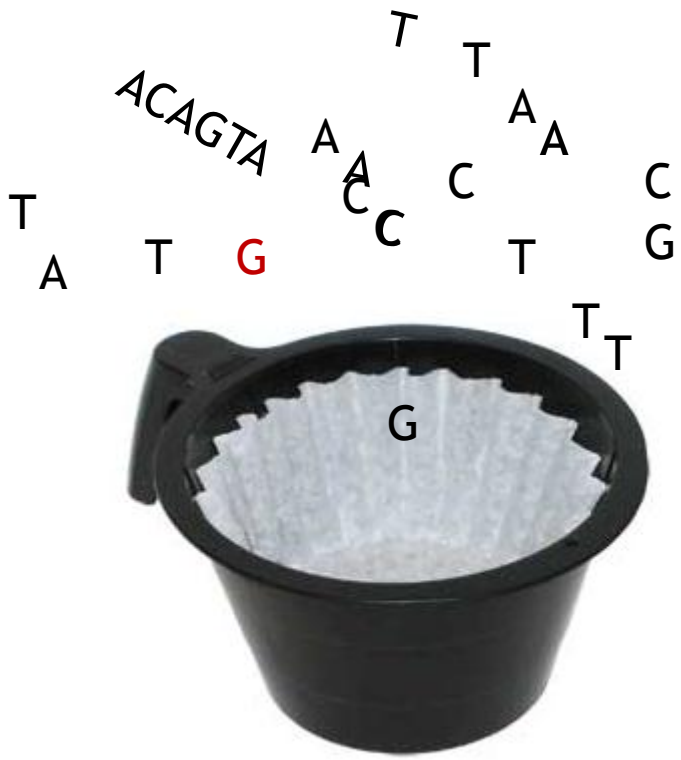
Abdou ElSharawy, Michael Forster, Nadine Schracke, Andreas Keller, Ingo Thomsen, Britt-Sabina Petersen, Björn Stade, Peer Stähler, Stefan Schreiber, Philip Rosenstiel and **Andre Franke**

For all author emails, please log on.

BMC Genomics 2012, **13**:417 doi:10.1186/1471-2164-13-417
Published: 22 August 2012

http://www.ikmb.uni-kiel.de/pibase/

Challenge: Annotate, Filter & Prioritize !



ikmb
Institut für Klinische Molekularbiologie

Omicia

Home | About | Sign-In

Unlocking Individualized Medicine

Advances in whole-genome sequencing technology are paving the way for genome analysis to become a routine part of healthcare delivery. Interpretation of an individual's genome sequence is now the key factor limiting the utility of that data for clinical applications. Omicia addresses this analytical bottleneck by helping researchers and clinicians better understand and interpret individual genetic variations, translating genomic insights into improved patient care.

Get Started Now

Introducing Omicia Opal

Omicia is pleased to announce the availability of Opal, a secure informatics platform that enables researchers to analyze genomes and prioritize disease-causing variants and genes. Opal combines powerful, peer-reviewed analysis tools with proprietary disease gene sets into an interactive genome mining, filtering, prioritizing, and reporting environment.

Opal also gives you access to the award winning VAAST algorithm (the Variant Annotation, Analysis and Selection Tool), a probabilistic search tool for identifying damaged genes and their disease-causing variants in personal genome sequences.

Sign-up and start interpreting your genomes today!

http://www.omicia.com/index.html

ikmb
Institut für Klinische Molekularbiologie

ANNOVAR

- Home
- Download
- Quick Start-up Guide
- Prepare Database
- Prepare Input File
- **Annotation**
- Gene-based
- Region-based
- Filter-based
- Accessary Programs
- FAQ

ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

1. **Gene-based annotation**: identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected. Users can flexibly use RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, or many other gene definition systems.
2. **Region-based annotations:** identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNAse I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.
3. **Filter-based annotation:** identify variants that are reported in dbSNP, or identify the subset of common SNPs (MAF>1%) in the 1000 Genome Project, or identify subset of non-synonymous SNPs with SIFT score>0.05, or find intergenic variants with GERP++ score>2, or many other annotations on specific mutations.
4. **Other functionalities:** Retrieve the nucleotide sequence in any user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, and other utilities.

SUMMARIZE_ANNOVAR is a script within the ANNOVAR package that is very popular among users. Given a list of variants from whole-exome or whole-genome sequencing, it will generate an Excel-compatible file with gene annotation, amino acid change annotation, SIFT scores, PolyPhen scores, LRT scores, MutationTaster scores, PhyloP conservation scores, GERP++ conservation scores, dbSNP identifiers, 1000 Genomes Project allele frequencies, NHLBI-ESP 5400 exome project allele frequencies and other information.

In a modern desktop computer (3GHz Intel Xeon CPU, 8Gb memory), for 4.7 million variants, ANNOVAR requires ~4 minutes to perform gene-based functional annotation, or ~15 minutes to perform stepwise "variants reduction" procedure, making it practical to handle hundreds of human genomes in a day.

What's new:

NEW! **2012Nov04**: The **NHLBI 6500 Exome data sets with indels and chrY calls** is available from ANNOVAR now! Use keyword esp6500si_ea, esp6500si_aa and esp6500si_all to download

http://www.openbioinformatics.org/annovar/

SeattleSeq Annotation 137

Sponsored by SeattleSNPs and SeattleSeq

- About SeattleSeq Annotation
- How to Use
- Build Notes
- Download Example Input Files
- SeattleSeq Annot. for hg18
- Contact Us

Input Variation List File for Annotation (NCBI 37 / hg19)

enter e-mail address:

Datei auswählen Keine Datei ausgewählt

input file format:
(SNVs only unless otherwise indicated)

- Maq
- GFF
- CASAVA
- VCF *(SNVs only)*
  specify output file format:
  - SeattleSeq Annotation file format
  - VCF file format
- custom
- one genotype per line
- GATK bed *(indels only)*
- VCF *(indels only)*
- VCF SNVs and Indels *(both)*

add more annotation:

SeattleSeq Annotation was most recently updated November 14, 2012. The current version is 8.00.

This site, based on dbSNP 137 is now active. See the build notes. However, response will be slow for a few weeks until we populate a database of cached annotations.

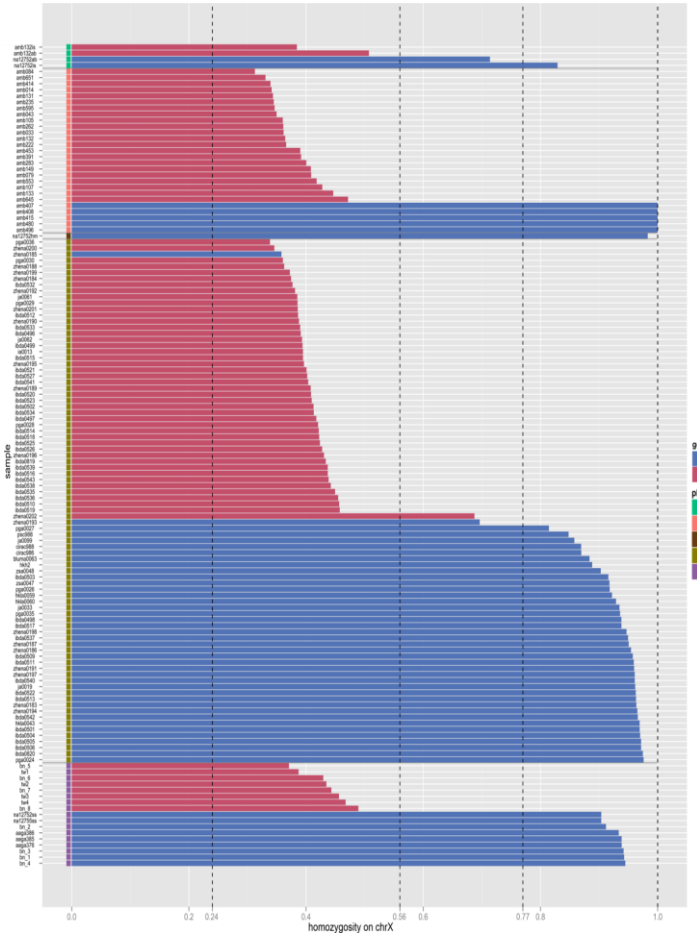
Thursday, December 06, 2012

ikmb
Institut für Klinische Molekularbiologie

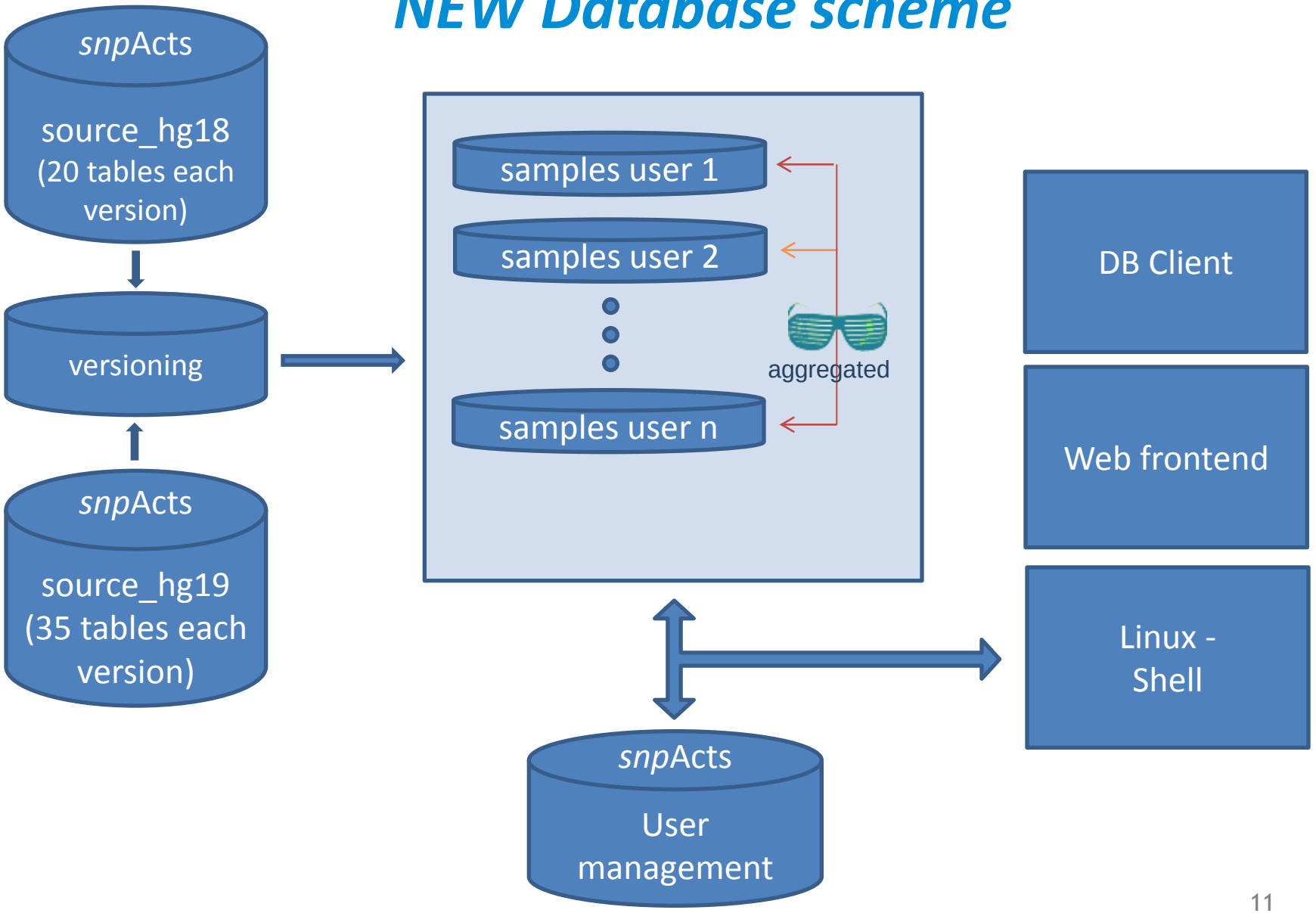
snpActs

analysis & categorisation tools

- Annotation of exome in app. 10 minutes
- Concordance check
- Gender check
- Mendel check
- Set operations & Venn diagrams
- Filter tools
- User-defined locus/variant sets
- Incl. other “lists” (PharmGKB etc.)
- Incl. pre-computed results for PolyPhen etc.
- IBD with SimWalk 2 for larger pedigrees

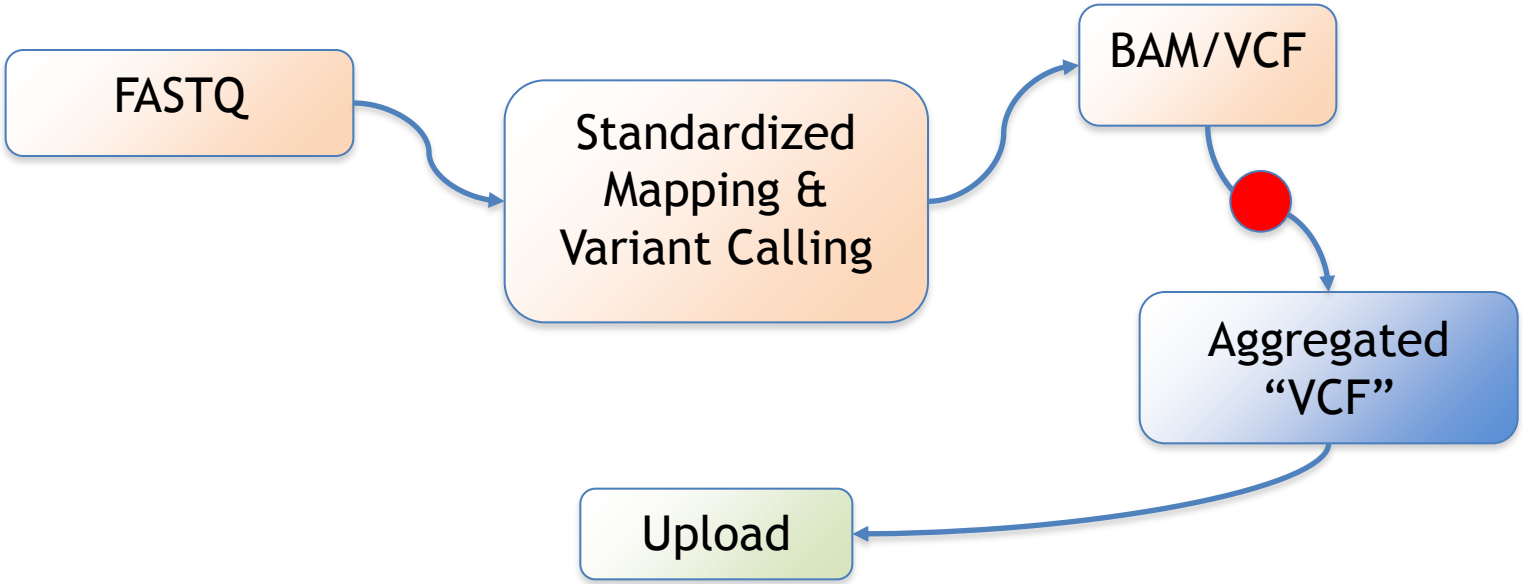


NEW Database scheme



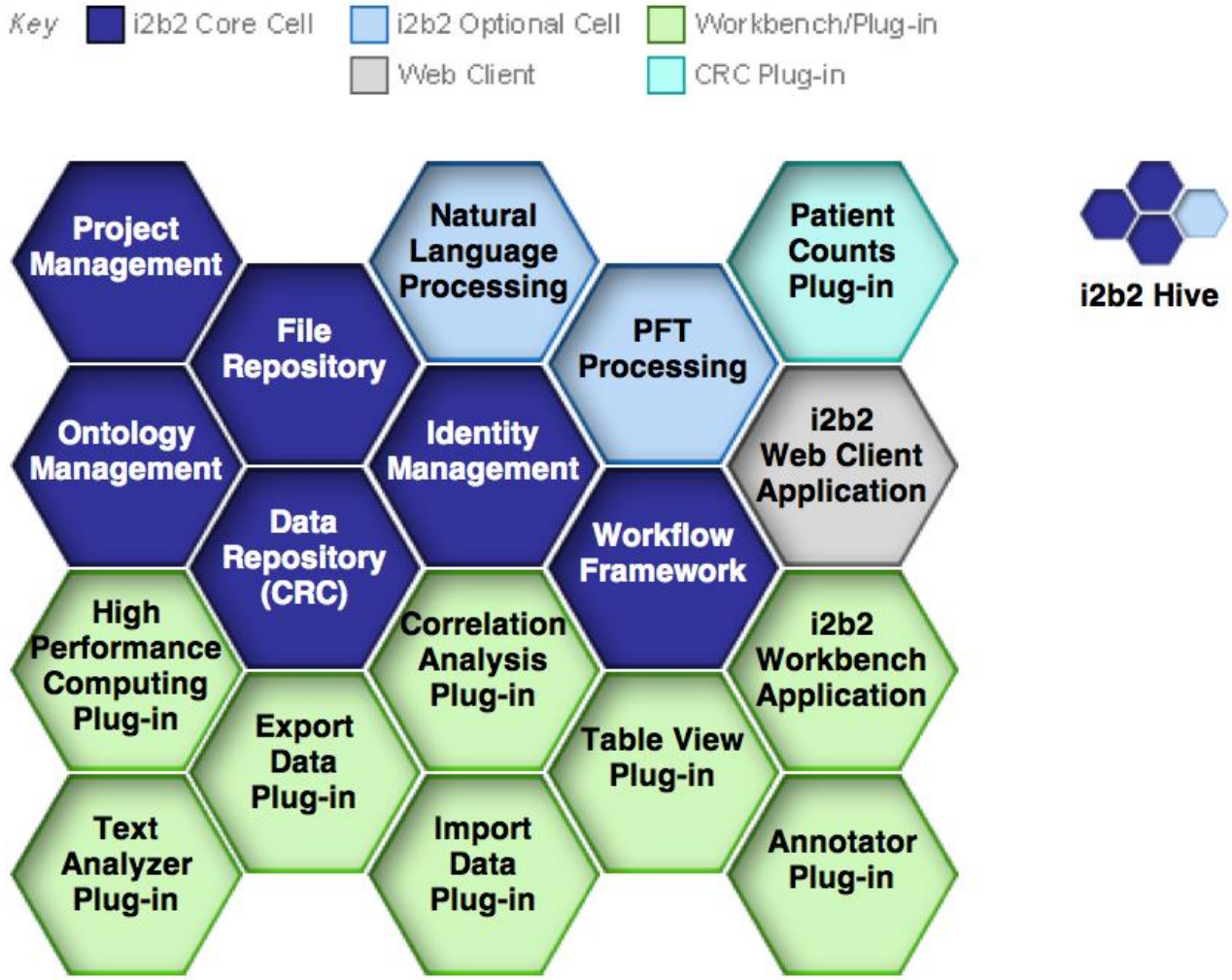
German SNV Database

- Is my variant private, rare or common in Germans? → frequency!
- See who produced the data.
- See the phenotype, if made available.
- Technical specs.

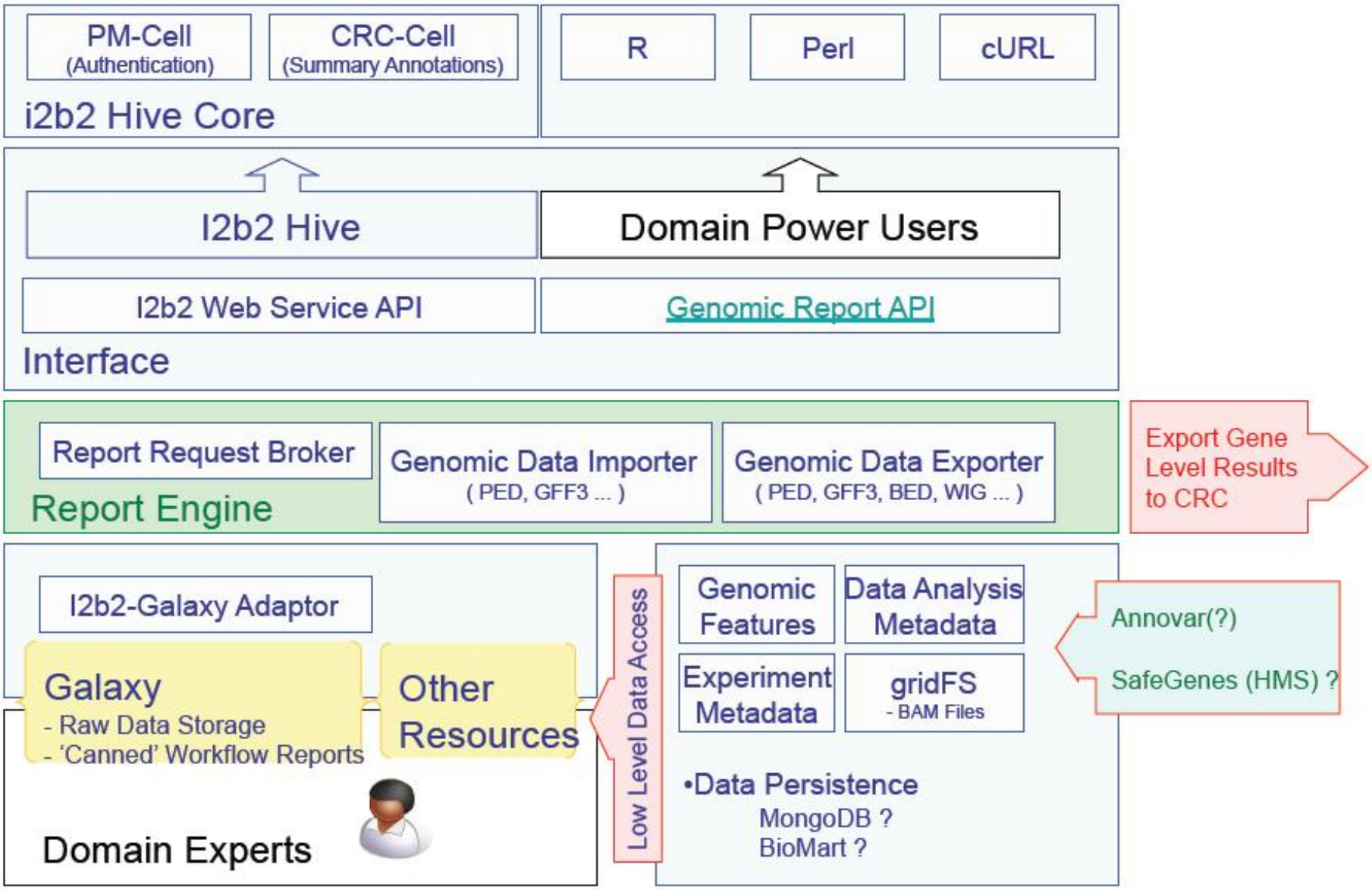


i2b2

Informatics for Integrating Biology & the Bedside



Component Diagram



GeneInsight®
for better care

Support Contact **Request a Demo**

About GeneInsight ▾ | GeneInsight Lab | GeneInsight Clinic | GeneInsight Network

TruSight + GeneInsight

The leaders in next-generation sequencing and clinical genetics interpretation and reporting are joining forces.

LEARN MORE

‹ ›

●○○○

The future is insight

GeneInsight Suite®, an IT platform developed at Partners HealthCare, streamlines interpretation and management of vast amounts of data, offering a key step towards the promise of personalized medicine and better patient care.

The future is indeed insight. GeneInsight.

We're Hiring!

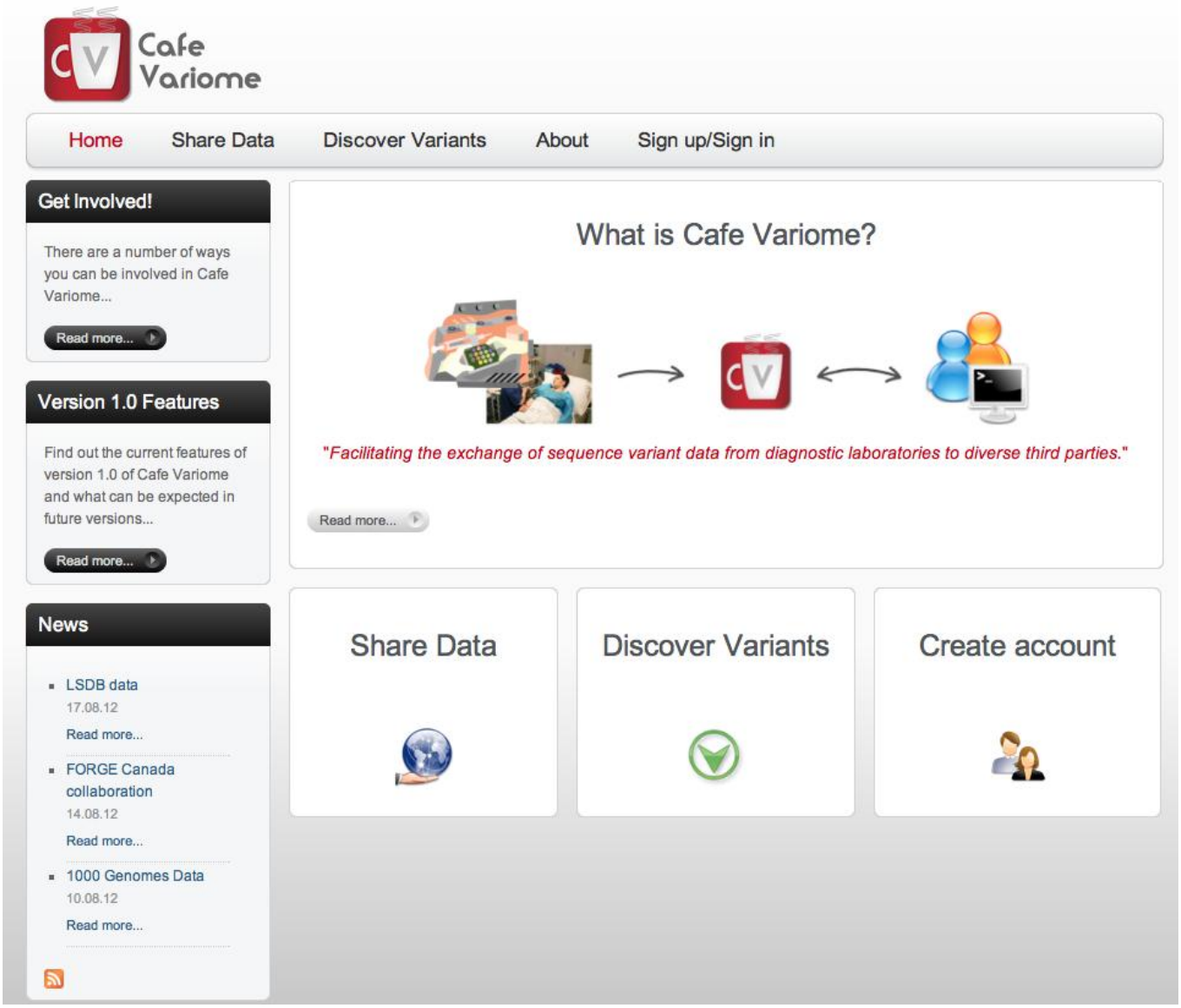
GeneInsight is seeking

News and Publications

Usability of a Novel Clinician Interface for

Upcoming GeneInsight Events

Personalized Medicine Conference, PCPGM


Cafe Variome
Home
Share Data
Discover Variants
About
Sign up/Sign in
Get Involved!
There are a number of ways you can be involved in Cafe Variome...
Read more...
Version 1.0 Features
Find out the current features of version 1.0 of Cafe Variome and what can be expected in future versions...
Read more...
News
LSDB data
17.08.12
Read more...
FORGE Canada collaboration
14.08.12
Read more...
1000 Genomes Data
10.08.12
Read more...
What is Cafe Variome?
"Facilitating the exchange of sequence variant data from diagnostic laboratories to diverse third parties."
Read more...
Share Data
Discover Variants
Create account

What do we need?

- Create a data warehouse for…
- Set up mining tools (see VAAST)…
- Allow more flexible and professional usage…
- Include various steps for quality control…
- Give users the chance to comment and rate… 

- Connect to i2b2!

SciEngines
massively parallel computing

SNP special interest group

SNP-SIG: Identification and annotation of SNPs in the context of structure, function, and disease.

GENERAL INFO:

WHAT: A one-day special interest group meeting
WHEN: July 14th, 2012
WHERE: ISMB 2012 venue in Long Beach (CA), USA.

SNP-SIG 2012 Meeting Programme and Abstracts PDF

SIG AIMS:

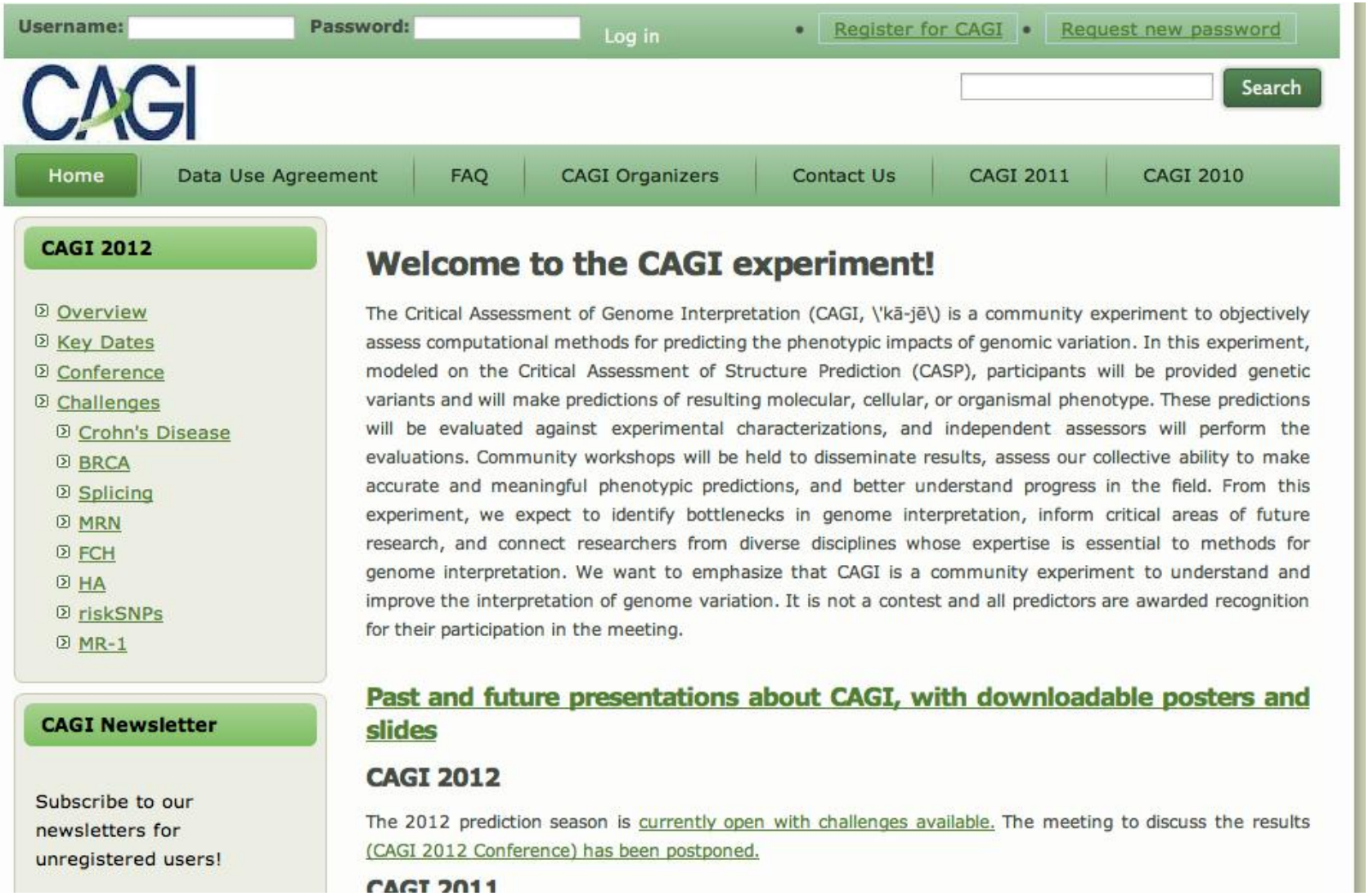
The primary goal of the SNP-SIG is to outline and discuss the recent advances in the **methodology for the annotation and analysis of genomic variation data**.

Building upon the experience of the **SNP-SIG 2011** in Vienna and other international workshops and meetings (e.g. **AIMM2010**, **CAGI**, **HGVS 2010** and **PSB2011**) the SNP-SIG will serve to **build a research network** , facilitating the exchange of ideas and the establishment of new collaborations within the community. Thus, SNP-SIG will strive to meaningfully contribute to the management of the complexity of the analysis and evaluation of genetic variation.

We are interested in attracting submissions describing original work in all the fields of genomic variation research including, but not limited to "genomic variation in":

- sequence analysis
- protein structure and function
- protein interactions and molecular networks
- transcriptomics and gene regulation
- disease models and epidemiology
- population genomics and evolution
- comparative genomics

CAGI



https://genomeinterpretation.org

Facts & Figures of the Group

- Academic sequencing center

  → broad spectrum of applications and includes DX!
- 3.5 people for Hardware / 4 for Software / 2.5 Scientists / 4 PhD
- Investments into dedicated IT infrastructure in millions of €
- For most purposes we use:
  - FASTQC, BWA, SAMTools, BEDTools, GATK, Picard, …

Inflammation at Interfaces

NGFN Nationales Genomforschungsnetz

READNA

ESGI