Max-Planck-Institut für Psychiatrie München

Internationale Aktivitäten zum Variantendatenbanking

Berlin 07 Dez 2012

Thomas Bettecken

Das Problem

Max-Planck-Institut für Psychiatrie

Varianten im menschlichen Genom

53.558.214 (38.077.993)
Varianten mit rs-Nr. (validierte)

22.508.883
davon in Genen

(exemplarisch - dbSNP build 137 26 Juni 2012)

Es gibt noch viel mehr Varianten !

Hat eine Variante eine Bedeutung im Hinblick auf Gesundheit und Krankheit ?

Welche Bedeutung?

Nach Mendel vererbte Erkrankungen I

Max-Planck-Institut für Psychiatrie

Autosomal rezessiv

Cystische Fibrose / CF Häufigkeit 1:2.500
Mutationen im CFTR-Gen
Heterozygotenfrequenz ~ 1:25 (~4%)
d.h. alle > 1.600 Mutationen zusammen
Häufigste Mutation delta-F-508
Allelfrequenz ~3% in Zentraleuropa
Allelfrequenz (Nicht-delta-F-508-Mutation) < 0.1%

Autosomal dominant

Polycystische Nierenerkrankung / ADPKD
Häufigkeit 1:400 – 1:1.000
~85% Mutationen in PKD1 / Polycystin-1 (Chr. 16)
häufigste einzelne Mutation: PKD1 c.5014-5015delAG
weltweit 23 x in nicht-verwandten Patienten beschrieben
plus 436 weitere verschiedene Mutationen (Dez 2011)
Allelfrequenz (c.5014-5015delAG) in ADPKD-Pat: <2%
Allelfrequenz in einer Population 0.002% - 0.005%
Andere Mutation: Allelfrequenz in einer Population < 0.0002-0.0005%

Nach Mendel vererbte Erkrankungen II

Max-Planck-Institut für Psychiatrie

X-chromosomal rezessiv

Muskeldystrophie Duchenne / DMD Häufigkeit 1:3.500
(männliche Neugeborene)

große Deletionen und Duplikationen (80%)

Punkt-Mutationen / PM (20%)

(die meisten sind private Neumutationen)

Allelfrequenz (PM-DMD) in einer Population << 0.01 %

Fazit::

Die Häufigkeit von Mutationen für monogene Erkrankungen ist bestenfalls ~3%. Die allermeisten Mutationen sind jedoch viel, viel seltener.

Datenbanken für Genom-Varianten

dbSNP	Katalog, z.Zt. 53.558.214 „SNPs“ (bld 137),Annotation der Pathologie ursprünglich nicht vorgesehen.
1000 Genomes Project	Daten: Whole Genome Sequenzierung von 1092 meist gesunden (?) Individuen.
HGMD	Human Gene Mutation Database, sehr gute Annotation klinischer Daten. Link zu LSDBs.
EVS (Exome Variant Server)	NHLBI Exome Sequencing Project (ESP) Daten: Whole Exome Sequenzierungen von 6503 Proben (Herz-, Lungen-, Bluterkrankungen)
ClinVar / OMIM Alleles	Sehr zuverlässig - gut dokumentiert (publizierte Fälle), aber nicht vollständig.
GEN2PHEN / G2P	„Integrated Genetic Variation Catalogue“: Diabetes - Obesity - Heart Disease - Cancer

…

NCBI dbSNP

Max-Planck-Institut für Psychiatrie

Varianten im menschlichen Genom

53.558.214 (38.077.993)
Varianten mit rs-Nr. (validierte)

22.508.883
davon in Genen

(dbSNP build 137 26 Juni 2012)

Es gibt noch viel mehr Varianten !

dbSNP war geplant als Katalog der Einzelnukleotid-Polymorphismen im menschlichen Genom. Die Annotation der Pathologie der eingetragenen Varianten war initial nicht vorgesehen, ist inzwischen zwar kryptisch vorhanden, insgesamt aber sehr lückenhaft.

1000 Genomes Project http://www.1000genomes.org/

Max-Planck-Institut für Psychiatrie

„The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in the populations studied.“

„The samples for the 1000 Genomes Project mostly are anonymous and have no associated medical or phenotype data; for some of the populations the collectors have phenotype data but these data are not at Coriell and are not distributed.“

1.092 samples (from 14 populations) completed, Phase 1 (2.500 planned).

Depth of coverage is varying (low, medium, …). Estimated power to detect genomic SNPs of frequency of 1% is 99.3%, of frequency 0.1% is 70%.

Data is available as raw reads (fastq), aligned reads (bam) and in variant call format (vcf) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/**).**

The 1000 Genomes Project Consortium, Nature (2010) 467:1061-1073 („Pilot paper“) and Nature (2012) 491:56-65 („Phase 1“)

Max-Planck-Institut für Psychiatrie

Data is publicly available as raw reads (fastq), aligned reads (bam) and after SNP calling im variant call format (vcf) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/)

Example line for the vcf format:

****

#CHROM POS ID REF ALT QUAL FILTER
16 11164547 rs114843098 T G 100 PASS

INFO
ERATE=0.0021;AVGPOST=0.9977;AA=T;AN=2184;VT=SNP;AC=14;RSQ=0.8702;LDAF=0.0074;
SNPSOURCE=LOWCOV;THETA=0.0007;AF=0.01;AMR_AF=0.01;AFR_AF=0.02;EUR_AF=0.0013

FORMAT SAMPLE01 SAMPLE02 SAMPLE03 ...
GT:DS:GL 0|0:0.000:-0.00,-3.47,-5.00 1|0:1.100:-5.00,-1.17,-0.03 0|1:1.000:-3.585,-0.00229916,-2.2993 ...

****

Legend: (GT - Genotype, DS - Genotype Dosage, GL - Genotype Likelihood)

1KGP useful: For lookup whether a „new“ variant has been seen before, estimating allele/genotype frequencies. Individual genotypes. For more ?

HGMD – Human Gene Mutation Database www.hgmd.cf.ac.uk

Max-Planck-Institut für Psychiatrie

The **Human Gene Mutation Database (HGMD)** represents an attempt to collate known (published) gene lesions responsible for human inherited disease. This database, whilst originally established for the study of mutational mechanisms in human genes (Cooper and Krawczak 1993), has now acquired a much broader utility in that it embodies an up-to-date and comprehensive reference source to the spectrum of inherited human gene lesions. Thus, **HGMD** provides information of practical diagnostic importance to (i) researchers and diagnosticians in human molecular genetics, (ii) physicians interested in a particular inherited condition in a given patient or family, and (iii) genetic counsellors.

Public entries: 92.715 (as of Dec 2012)
Total entries: 103.522 (difference will be made public within 3 years)

For entry of a variant, a public documentation of the phenotype is required.
Link to 349 Locus Specific Databases (LSDB).

Stenson et al (2003), The Human Gene Mutation Database (HGMD®): 2003 Update.
Hum Mutat (2003) 21:577-581.

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server (EVS) (evs.gs.washington.edu/EVS/)

Max-Planck-Institut für Psychiatrie

Whole Exome-Sequenzierungen genomischer DNA
(alles Patienten mit „Herz- Lungen- und Bluterkrankungen“)

Analyse eines Subsets von 2440 Individuen
(Tennessen et al., Science 2012, 337:64-69, Juli 2012)

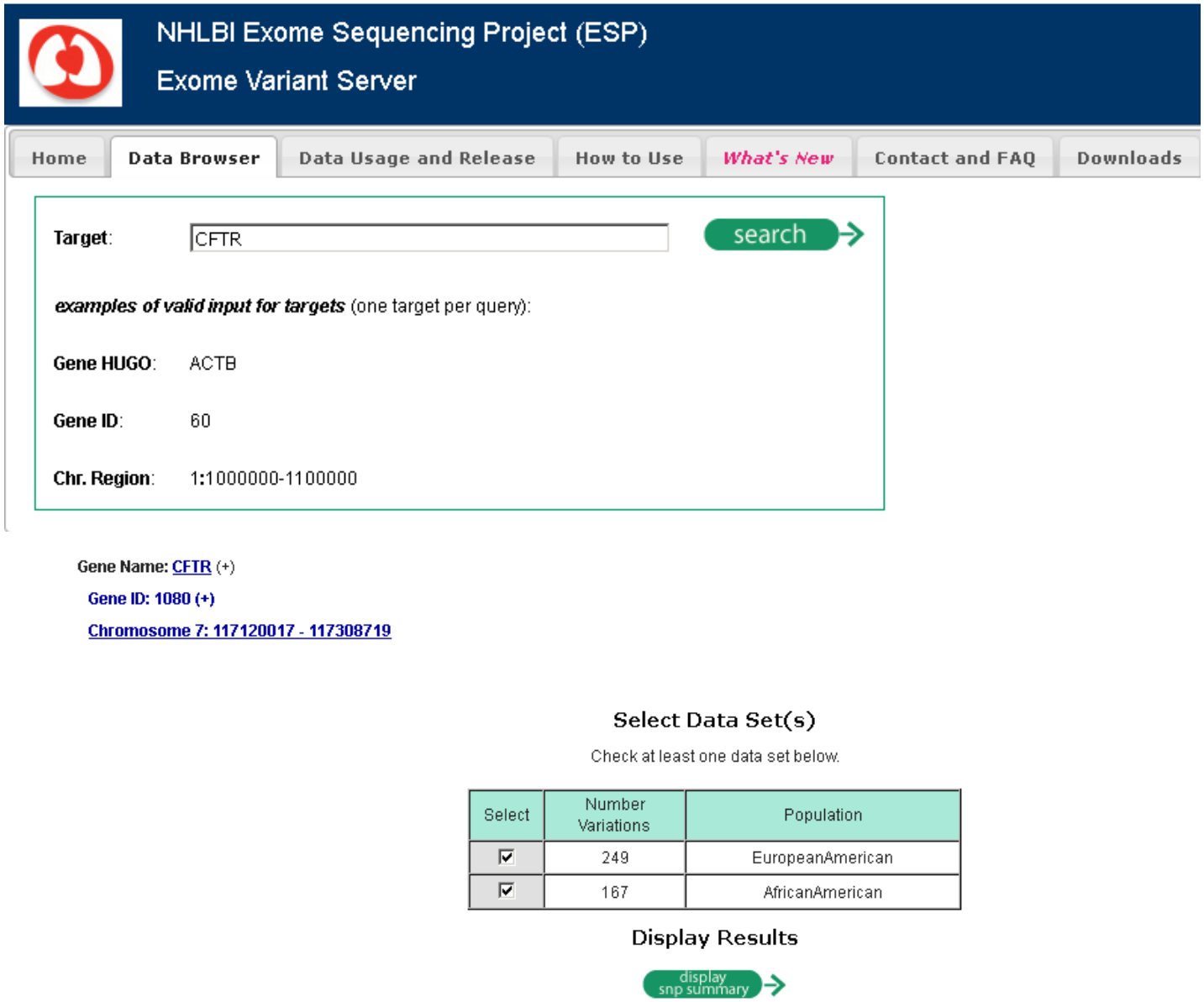
Sequenziert wurden 15.585 Protein-Codierende Gene (3 Exome Definitionen: CCDS 2008 26 MB, Roche EZ Cap v1 32 MB, Roche EZ Cap v2 34 MB)

Gefunden wurden >500.000 Varianten
davon 82% zuvor unbekannt
davon 86% mit MAF < 0.5%
Jedes Individuum trägt im Durchschnitt: 13.595 SNPs
2.3% dieser Varianten im Genom eines Individuums betreffen/ändern die Funktion von ~313 Genen

Server: Whole Exome-Daten von 6503 Patienten (ESP6500SI, 4300 AE – Americans of European descent, 2203 AA – Americans of African Descent)

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server (EVS) (evs.gs.washington.edu/EVS/)

Max-Planck-Institut für Psychiatrie



NHLBI Exome Sequencing Project (ESP)
Exome Variant Server (EVS) (evs.gs.washington.edu/EVS/)

Max-Planck-Institut für Psychiatrie

Gene Name: CFTR **(Gene ID: 1080)** (+)

Chromosome 7: 117120017 - 117308719

Population: EuropeanAmerican

GWAS Catalog: CFTR

KEGG Pathway: CFTR

Sanger COSMIC: CFTR

PPI STRING 9.0: CFTR

Variation Color Code:
splice or nonsense or frameshift
missense
coding-synonymous
coding
utr
codingComplex

Dow

File F

Zip F

Add or Remove Columns (Description of Columns)

☑ dbSNP rs ID	☑ Alleles	☑ EA Allele Count	☑ AA Allele Count	☐ Allele Count	☑ Sample Read Depth	☐
☐ Genes	☑ Gene Accession #	☑ GVS Function	☑ Amino Acid	☑ Protein Position	☑ cDNA Position	☐
☐ Chimp Allele	☐ Conservation (phastCons)	☐ Conservation (GERP)	☐ Grantham Score	☐ PolyPhen Prediction	☑ Clinical Link	☐
☐ EA Genotype Count	☐ AA Genotype Count	☐ Genotype Count	☐ Illumina HumanExome Chip		☐ GWAS Hits	

Sort Variants by Variant Pos reset

NHLBI Exome Sequencing Project (ESP)
Exome Variant Server (EVS) (evs.gs.washington.edu/EVS/)

Variant Pos	rs ID	A1/R Alleles	EA Allele #	AA Allele #	Avg. Sample Read Depth	mRNA Accession #	GVS Function	Amino Acid	Protein Pos.	cDNA Pos.	Clinical Link
7:117188755	unknown	A/G	A=2/G=8560	A=0/G=4404	24	NM_000492.3	missense	SER,GLY	424/1481	1270	unknown
7:117188810	unknown	**A1/R**	A1=1/R=8235	A1=0/R=4256	27	NM_000492.3	frameshift	none	NA	NA	unknown
7:117188812	rs147422190	T/G	T=2/G=8586	T=2/G=4398	27	NM_000492.3	missense	TYR,ASP	443/1481	1327	unknown
7:117188814	rs148056476	G/T	G=1/T=8587	G=0/T=4400	27	NM_000492.3	missense	GLU,ASP	443/1481	1329	unknown
7:117188823	unknown	**A1/R**	A1=2/R=8246	A1=1/R=4261	27	NM_000492.3	frameshift	none	NA	NA	unknown
7:117188849	unknown	A/C	A=1/C=8599	A=0/C=4406	31	NM_000492.3	missense	GLU,ALA	455/1481	1364	unknown
7:117188850	rs79074685	A/G	A=1/G=8599	A=16/G=4390	32	NM_000492.3	coding-synonymous	none	455/1481	1365	unknown
7:117188919	unknown	C/T	C=3/T=8593	C=0/T=4400	23	NM_000492.3	intron	none	NA	NA	unknown
7:117188928	unknown	**A1/A2/R**	A1=105/A2=92 /R=8029	A1=54/A2=45/R=4121	18	NM_000492.3	intron	none	NA	NA	unknown
7:117199524	rs1800089	T/C	T=5/C=8595	T=0/C=4406	106	NM_000492.3	missense	PHE,LEU	467/1481	1399	unknown
7:117199525	rs139573311	C/T	C=2/T=8598	C=1/T=4405	106	NM_000492.3	missense	PRO,LEU	467/1481	1400	unknown
7:117199532	rs143218779	T/G	T=1/G=8599	T=0/G=4406	108	NM_000492.3	missense	ILE,MET	469/1481	1407	unknown
7:117199533	rs213950	A/G	A=4026/G=4574	A=3747/G=659	108	NM_000492.3	missense	MET,VAL	470/1481	1408	omim_link
7:117199579	rs143980575	C/G	C=1/G=8599	C=1/G=4405	119	NM_000492.3	missense	THR,SER	485/1481	1454	unknown
7:117199602	rs77101217	T/C	T=2/C=8598	T=0/C=4406	111	NM_000492.3	stop-gained	stop,GLN	493/1481	1477	omim_link

CFTR Variation Viewer Download report (283745 bytes)

Gene CFTR; cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)

Description ATP-binding cassette sub-family C member 7 | ATP-binding cassette transporter sub-family C member 7 | cAMP-dependent chloride channel | channel conductance-controlling ATPase | cystic fibrosis transmembrane conductance regulator
Also known as: tcag7.78, ABC35, ABCC7, CF, CFTR/MRP, MRP7, TNR-CFTR, dJ760C5.1

Species *Homo sapiens*

Gene Reference Sequences NG_016465.1 genomic, NM_000492.3 transcript, NP_000483.3 protein
variation locations are

Links HGMD, Panther, Gene,

Observed Variation | Page 6 of 23 | Displaying results 101 - 120 of 458

Var Class	Genomic	Transcript	Protein	Clinical interpretation	Test status	Dep	Obs	Fre
SNC	g.84509T>C	c.1400T>C	p.Leu467Pro	probable-pathogenic		2	2	
SNC	g.84517G>A	c.1408G>A	p.Val470Met	no known pathogenicity		8	63	
SNC	g.84547G>C			pathogenic		4	6	
SNC	g.84547G>T	c.1438G>T	p.Gly480Cys	pathogenic		4	6	
SNC	g.84547G>A			pathogenic		4	6	
SNC	g.84584C>T	c.1475C>T	p.Ser492Phe	pathogenic		1	1	
SNC	g.84586C>T	c.1477C>T	p.Gln493Ter	pathogenic		1	3	

OMIM Variant Viewer, CFTR entry:

http://www.ncbi.nlm.nih.gov/sites/varvu?gene=1080&rs=213950|
http://omim.org/entry/602421#0023

Max-Planck-Institut für Psychiatrie

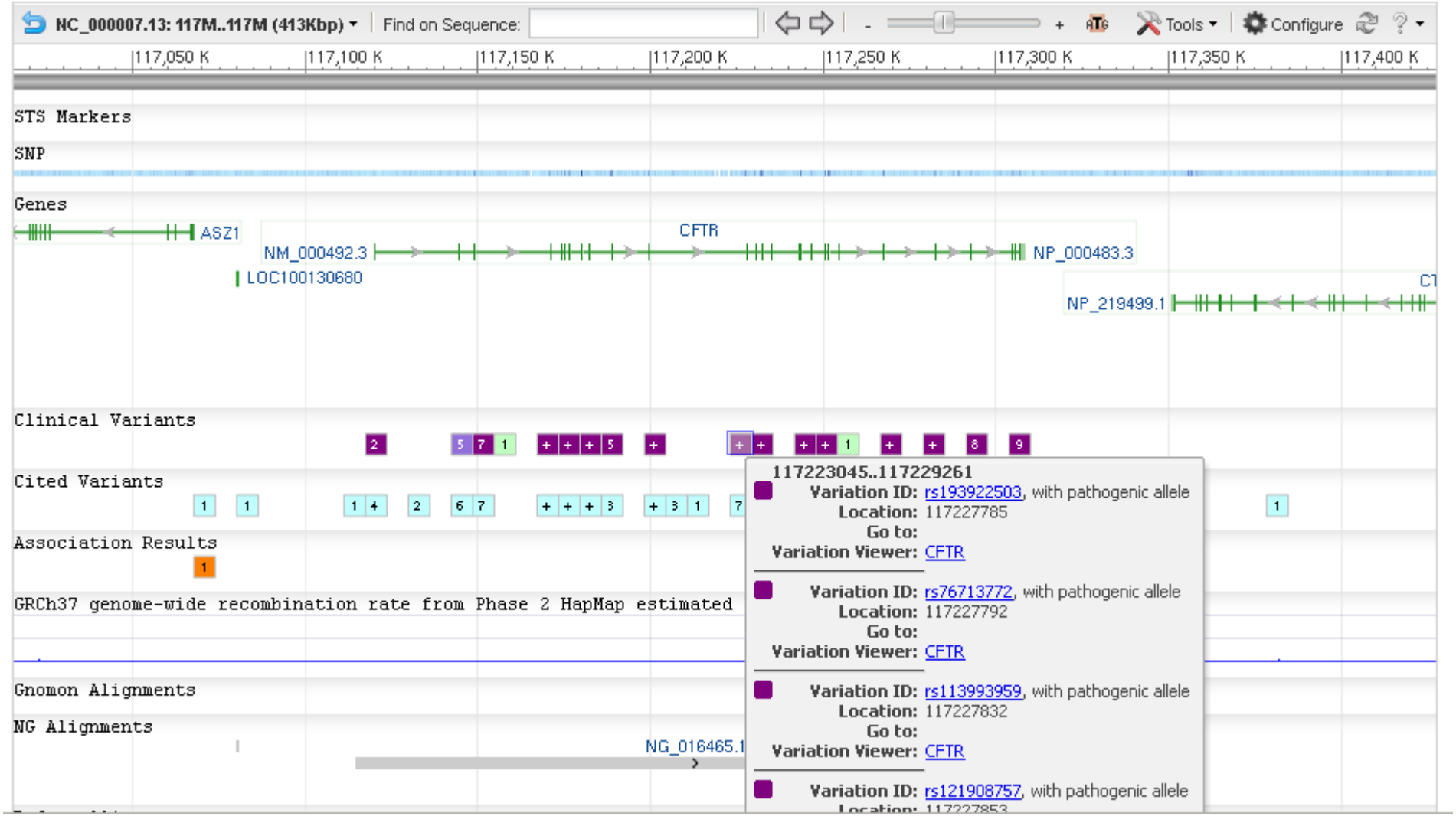
What is ClinVar? (see also supplementary info)

„The goal of ClinVar is to provide a freely accessible, public archive of reports of the relationships among human variations and phenotypes along with supporting evidence. By so doing, ClinVar will facilitate access to and communication about the relationships asserted between human variation and observed health status. ClinVar collects reports of variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. The alleles described in the submissions are mapped to reference sequences, and reported according to the HGVS standard. ClinVar then presents the data for individual users, laboratories that want to incorporate it in their daily workflow, and organizations that want to incorporate it into their own applications.

We emphasize reporting structured evidence supporting any genotype-phenotype relationship, in order to support computational (re)evaluation, both of genotypes and assertions, enabling the ongoing evolution and development of knowledge regarding variations and associated phenotypes.“

ClinVar Variation Reporter

Max-Planck-Institut für Psychiatrie



Linking to „CFTR“ leads to the same NCBI CFTR Variation viewer as above (458 entries)

ClinVar www.ncbi.nlm.nih.gov/clinvar/

Max-Planck-Institut für Psychiatrie

Principle Values:

1. **Clinical grade mutation database**
    a. Evidence towards or against pathogenicity
2. **Make data freely available**
3. **pre-competitive space**
4. **expectations of reciprocity**
    a. Users to also contribute with an obligation to submit. Create tools to monitor, allow access?
5. **Neutral party to host**
    a. New non-profit organization entity that can raise money?
6. **Criteria of threshold for submission**
    a. Verification of variant (only by CLIA certified labs or Research labs meeting criteria?)
    b. Tiers of submission
7. **Publication:**
    a. Microattribution – submission be accompanied by submitter ID.
8. **Consent model**:
    a. Opt out mechanism of data acquisition.
        i. Requirements of informing physicians, ISCA website model, pre-reviewed with Jim Ostell, NCBI – OHRPP get through IRB not formally
    b. IRB requirements
    c. Prior data – retrospective prior to initiation to opt-out allowed to send de-identified and phenotypic info but not raw data files.
    d. Models for opt-out for every patient in hospital.

For more info see: ClinVar Policy Notes and ClinVar Standards Notes.

Max-Planck-Institut für Psychiatrie

„The GEN2PHEN project aims to unify human and model organism genetic variation databases towards increasingly holistic views into Genotype-To-Phenotype (G2P) data, and to link this system into other biomedical knowledge sources via genome browser functionality.“

Partner: 17 Europäische Gruppen, davon aus D nur EMBL, plus 1 Indische Gruppe, plus 1 Südafrikanische Gruppe.

Max-Planck-Institut für Psychiatrie

Strategy and Aim

The **GEN2PHEN** project has the overall ambition of unifying **human and model organism genetic variation databases**, and doing this in such a way that the resulting holistic view of G2P data can be blended with all other biomedical database domains via one or more central genome browsers. The project will put in place the main building blocks needed to move substantially from today's G2P database situation towards the ultimate future of a complete biomedical knowledge environment. **The project will then utilise these building blocks to construct a first-generation version of a G2P knowledge environment by the project's end. This will consist of a European-centred but globally networked hierarchy of bioinformatics GRID-linked databases, tools and standards, all tied into the Ensembl genome browser.** To ensure the project builds something that truly works and tangibly benefits the community, rather than merely devising potentially useful technologies, we have focussed the project's objectives on the three essential components of a functioning G2P database system. These can be viewed as three legs of a 'stool', each of which must be robust for the stool to properly function (see Figure 1).

Max-Planck-Institut für Psychiatrie

Cafe Variome Overview and Core Concept

Diagnostics laboratories assess DNA samples from many patients with various inherited disorders, and so produce a great wealth of data on the genetic basis of disease. Unfortunately, those data are not usually shared with others. To address this gross deficiency, we are constructing a system that will facilitate the automated transfer of diagnostic laboratory data to the wider community, via an Internet-based Café for Routine Genetic data Exchange (Cafe Variome).

Diagnostic laboratories are not reluctant to release their data. Instead, the obstacles are merely practical: First, diagnostic laboratory personnel do not have time nor funding to manually submit data to Internet depositories such as Locus Specific databases (LSDBs). Second, diagnostic laboratories would receive no recognition or reward for releasing their data, giving them little incentive to even try.

The Cafe Variome approach takes account of the real-world obstacles and the needs of diverse LSDBs (insights provided by GEN2PHEN: http://www.gen2phen.org).

EU-Projekt Laufzeit Jan 2008 - Dez 2012 12 Mio Euro
scheint seit Ende 2010 weniger aktiv zu sein.

Datenbanken für Genom-Varianten
Zusammenfassung

Max-Planck-Institut für Psychiatrie

dbSNP	Katalog, z.Zt. 53.558.214 „SNPs“ (bld 137),Annotation der Pathologie ursprünglich nicht vorgesehen.
1000 Genomes Project	Daten: Whole Genome Sequenzierung von 1092 meist gesunden (?) Individuen.
HGMD	Human Gene Mutation Database, sehr gute Annotation klinischer Daten. Link zu LSDBs.
EVS (Exome Variant Server)	NHLBI Exome Sequencing Project (ESP) Daten: Whole Exome Sequenzierungen von 6503 Proben (Herz-, Lungen-, Bluterkrankungen)
ClinVar / OMIM Alleles	Sehr zuverlässig - gut dokumentiert (publizierte Fälle), aber nicht vollständig.
GEN2PHEN / G2P	„Integrated Genetic Variation Catalogue“: Diabetes - Obesity - Heart Disease - Cancer

…

Nachtrag (Diskussionsbemerkung)

Max-Planck-Institut für Psychiatrie

Zu GEN2PHEN

Peter Robinson bemerkt, daß das Projekt GEN2PHEN durchaus weiterhin aktiv ist und produktiv ist.

Nachtrag (nicht präsentiert am 7.12.12)

Max-Planck-Institut für Psychiatrie

LOVD **Leiden Open Variation Database** (www.lovd.nl)

Partially finanziert von der EU im Rahmen des GEN2PHEN Projekts.

„A flexible, free tool for gene-centered collection, curation and display of DNA variation“.

LOVD 2.0 wurde erfolgreich getestet mit mehr als 5.000 Genen, 1.000.000 Mutationen, 500.000 Patienten und 100.000 Einsendern.

Open Source. Kann heruntergeladen und installiert werden auf einem Linux-Server mit Apache, PHP und MySQL und 150 MB Plattenspeicher.

Es ist jedoch noch nicht klar, ob auch Exom-weite Daten effektiv verwaltet werden können.

Fokkema et al. Hum Mutat. 2011 May;32(5):557-63.

Nachtrag (nicht präsentiert am 7.12.12)

Max-Planck-Institut für Psychiatrie