Retrieving information on genes and proteins from biological and genomic databases

Marylyn D Ritchie, PhD
Professor, Biochemistry and Molecular Biology
Director, Center for Systems Genomics
The Pennsylvania State University
GenBank

- Repository of nucleic acid sequences
- As of 2001, held 9.5 billion bases in 8.2 million entries

<table>
<thead>
<tr>
<th>Release</th>
<th>Date</th>
<th>GenBank Bases</th>
<th>GenBank Sequences</th>
<th>WGS Bases</th>
<th>WGS Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Dec 1982</td>
<td>680338</td>
<td>606</td>
<td></td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>Aug 2000</td>
<td>9,545,724,824</td>
<td>8,214,339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>129</td>
<td>Apr 2002</td>
<td>19,072,679,701</td>
<td>16,769,983</td>
<td>692,266,338</td>
<td>172,768</td>
</tr>
</tbody>
</table>
GenBank
SCN5A – 32 exons
Homo sapiens sodium channel, voltage-gated, type V, alpha subunit (SCN5A), transcript variant 1, mRNA

NCBI Reference Sequence: NM_198056.2

Go to:

LOCUS NM_198056 8504 bp mRNA linear FRI 03-MAY-2014
DEFINITION Homo sapiens sodium channel, voltage-gated, type V, alpha subunit (SCN5A), transcript variant 1, mRNA.
ACCESSION NM_198056 NM_198056.1
VERSION NM_198056.2 GI:124518659
KEYWORDS RefSeq.
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
  Mammalia; Eutheria; Euarchontoglires; Primates; Hominoidea;
  Hominoidea; Homo.
REFERENCE 1 (bases 1 to 8504)
  Linke T, Bettle M, Perez-Ville F, Perez OJ, Scornik FS, Benndorf K,
  Pagans S, Zimmer T and Brugada R.
TITLE Protein arginine methyl transferases-3 and -5 increase cell surface
  expression of cardiac sodium channel
PUBLISHED 2013

<table>
<thead>
<tr>
<th>Amino Acid Sequence</th>
<th>Nucleotide Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><strong>/translation=</strong></td>
<td>1 agacggggcc ggcggccgta ggatgcaggg atgcctcccc ggggctgct gacgcttggc</td>
</tr>
<tr>
<td>GLPEEARPQDLQASKLPLDPQLGNPPQELIGELEDLDPFSTQK</td>
<td>61 ccagtgcccc gacgcgcccc gcggcagcag tccggccaa gcaagagcgg cccaccccc</td>
</tr>
<tr>
<td>RFRTSVALLQHPSLFLHQSLAARAKVLHSLFLNMLTLCNQVMFA</td>
<td>121 ggccggccgc ggggacaccg agttctcccac cagggcagtg gaggagagc cgtgcggcag</td>
</tr>
<tr>
<td>EYTFTAIYTFEFLVKLARGFCWHTFNLQDPWNLWDSFIVINAYMTT</td>
<td>181 agcaggagta ggaagtgcca aacctctctct tacctccaggg cgacagcagc cttggctagg</td>
</tr>
<tr>
<td>RRTFRVLNLWLGHTINQFVWLGIVSWSNQNLTHYIMVWYVIFSFLNS</td>
<td>241 tcaggccccg gttctggctca gcacagagca agcgccagct gaggacaggc ggagacggct</td>
</tr>
<tr>
<td>RHKVCNRTALNGTSWEDATMILWNLGKMTNLLNGTSDLCEA</td>
<td>301 caacaccctt gcagagacac gcggcagcgg tcgcggccag aagcctgcag cagctggcag</td>
</tr>
<tr>
<td>CPEGYRKLACNGPDGHTSFDSFWAFAELFRLMPTDLCVERLGYQTL</td>
<td>361 tcggactcgc gcagccctaa aagcctggcc atccatttgct caatccattg cagctggcag</td>
</tr>
<tr>
<td>FMLVFLGFSVLNWILAVAMAYEEQNXQTIAETEEKEKRQFQETM</td>
<td>421 tcggagaccc ctggagcagc gcggctggct gatttcgttg tctactagct tcctagatcg</td>
</tr>
<tr>
<td>RVGDVTSLRSSLSVPLSNPVLHRRSKRMKSSGEGEDDPELSER</td>
<td>481 tgaataaagcc ctgcgctggc atgcctgctc tctagagcct tcctagatcg</td>
</tr>
<tr>
<td>LSQRLGTSRMIKPSRSSGIFRTTTLGEDAEDTENNGSEPE</td>
<td>541 cttcagcccc catggccagc gcggctggct gatttcgttg tctactagct tcctagatcg</td>
</tr>
<tr>
<td>LRTRSAQQPSGTSPSAGHRGNKSTGNCVDSGNQGPGAG</td>
<td>601 tcatctatgt gcacatcccc acacoaccgc ttgctctagg cccagcccct cacagcctgg</td>
</tr>
<tr>
<td>EHPPDDTTTSPEPGPGQNLTSQAPCVDGFEEGPARQALASVSLT</td>
<td>661 gagccagctg tgtctgctgt cttcagcccc catggccagc gcggctggct gatttcgttg tctactagct tcctagatcg</td>
</tr>
<tr>
<td>CPPCWNRLAQRYLWIECPWMSIKQGVKLVMDPFDLTLTIMCIVL</td>
<td>721 tcggctgccg tcggctgccg acctctctct acctgctgctc tctagagcct tcctagatcg</td>
</tr>
<tr>
<td>MTSEFEELMQQNVLGNTFQITAQPQFOQGNWFDIS</td>
<td>781 tcggctgccg tcggctgccg acctctctct acctgctgctc tctagagcct tcctagatcg</td>
</tr>
<tr>
<td>SRMSNLVSRLSRFFLRVFKLAWPTLNTLKVIIQNSVGLANLNT</td>
<td>841 cagccagcccc catggccagc gcggctggct gatttcgttg tctactagct tcctagatcg</td>
</tr>
<tr>
<td>GMQLGFQNLNLQERSSGGGLRWMHDFHAFLFIRLGCWEVET</td>
<td>901 tgagagcatt cggctggctg gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>CLCVFLWVMLIGNVLWNLFLALLSSFADNTAPEDREMNNQL</td>
<td>961 tcagctttcc gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>VRKTLMQVLLSRLQGGLQQAPLAAGQLGQWGTVFVTY</td>
<td>1021 taagccacat gcgtgctgctgacctagccag gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>GEQPGQGTPGPDEPVPCVPVIAAESDTRQDEEDEENSLETGEESKQ</td>
<td>1081 cgagccagcc gcggctggctg acctagccag gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>PPDSRTJSQVSASSAEASASQSEASAEQGGEGPTPE</td>
<td>1141 tcagagacgc cagcctgcttg gttctcagct gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>TAELEDQPIDLPQVKDQEDPCFGEVCGCCPCCACAVTTQAPGKWN</td>
<td>1201 cgccagcttg cggctggctg acctagccag gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>SWFETIFVNLLQGSLALEDFEIRKIKVELQYoutube</td>
<td>1261 attctctctt gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>FKKYFTNACWCLDFLVDVSLIVSNALGFAMNQPKIIESLAR</td>
<td>1321 gcctttccag gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>RVNNVAALGAIPSNMVLNCLFLFILVSVMGNFAGKFGRCINQ</td>
<td>1381 tcagagacgc cagcctgcttg gttctcagct gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>NNKSCQESNLNGLYWTVKGNVFDNVGAYLALLQCVTFWM</td>
<td>1441 atgagagcag aacaaacccc aacctgctgg agacggaggaa agaaggaagag ctgccctcg</td>
</tr>
<tr>
<td>POQPWLNLYNMYIFONGIFSIFGSLGIFDDNFQIONSDFQGG</td>
<td>1501 agggcattga aatgtcaagag aagacagtgc aagccctcag ctagggcttg gttgatagc</td>
</tr>
<tr>
<td>YNAMHKKGLGSKPQPPIRIKNKYQGIIFDIVTKQVDVTIFMIFLICL</td>
<td>1561 ttgctgctgg ttgcctcctgc gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>SPEKINLSAKIWLFTRICLFWNSLNHFDFVFWV</td>
<td>1621 gcagagagcc cagcctgcttg gttctcagct gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td>IIQKYFFSPIFLVRILARIRGLIRLRILGAKRFLTLALMPSLA</td>
<td>1681 agttctgctgg aaggaagctt cttctcaggg tgggagagct agctggcctc</td>
</tr>
<tr>
<td>FIYISGFCEFMVAGFIDMFMVFQFPFQFFQITTQSDWGG</td>
<td>1741 cgagagaccc tgggagagct cttctcaggg tgggagagct agctggcctc</td>
</tr>
<tr>
<td>YCDPTLNSINSGRGDCGASPAGVILFTYIIISFLIVNMIAILE</td>
<td>1801 acctggctgg tgagagcagtt tttcgatgag atggagatgg cacagccggg gaggacgagg</td>
</tr>
<tr>
<td>PLSEDMMDFYIEWEKTDEQTIPQFIYSVLSDFALDELPRIKAN</td>
<td>1861 acctggctgg tgagagcagtt tttcgatgag atggagatgg cacagccggg gaggacgagg</td>
</tr>
<tr>
<td>VSGDRHLILFATKRVGELSGMPKIYQEMIPEISYEVSCQICRSRL</td>
<td>1921 cagtcctgcg aaccctgccg cttcagctgg cttctcaggg tgggagagct agctggcctc</td>
</tr>
<tr>
<td>EESVSAMVQFRARHRQLQSLHANFLFQQAGSGLSEADEPERGL</td>
<td>1981 actcctgctgg tttcagctgg cttcagctgg cttctcaggg tgggagagct agctggcctc</td>
</tr>
<tr>
<td>PLGGPSISSSTSSFPSPPSYSDVTRATSDLNQLVGRSGDHYSEDLADFF</td>
<td>2041 gcctggctgg ttgcctcctgc gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td></td>
<td>2101 tgaagctggc cttcagctgg cttcagctgg cttctcaggg tgggagagct agctggcctc</td>
</tr>
<tr>
<td></td>
<td>2161 gcagagagcc gcggagagct ttgcctcctgc gcagagagag ctgggtgagt cagctagtgc ggtgatgcgg</td>
</tr>
<tr>
<td></td>
<td>2221 tagagagcttg cttcagctgg cttcagctgg cttctcaggg tgggagagct agctggcctc</td>
</tr>
</tbody>
</table>
SCN5A  sodium channel, voltage-gated, type V, alpha subunit [Homo sapiens (human)]

Summary

Official Symbol: SCN5A  provided by HGNC
Official Full Name: sodium channel, voltage-gated, type V, alpha subunit  provided by HGNC
Primary source: HGNC: HGNC:10553
See related: Ensembl: ENSG0000013873; HPRD:02543; MIM:600163; Vega: OTTHUM000000156166
Gene type: protein coding
RefSeq status: REVIEWED
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; chordata; craniata; vertebrata; euteleostomi; mammalia; eutheria; euarchontoglires; primates; haplorrhini;
catatrich; hominidae; Homo
Also known as: HB1; HB2; HH1; V1; HBB; ICCD; LQT3; SSS; CCD2; CMD1E; CMPD2; PFHB1; Nav1.5
Summary: The protein encoded by this gene is an integral membrane protein and tetrodotoxin-resistant voltage-gated sodium channel subunit. This protein is found primarily in cardiac muscle and is responsible for the initial upstroke of the action potential in an electrocardiogram. Defects in this gene are a cause of long QT syndrome type 3 (LQT3), an autosomal dominant cardiac disease. Alternative splicing results in several transcript variants encoding different isoforms. [provided by RefSeq, Jul 2008]

Genomic context

Location: 3p21
Exon count: 32
### SNP linked to Gene (geneID:6331) Via Contig Annotation

The SNP GeneView page only reports human variation on GRCh38. A new Variation Viewer is available to view the gene SCN5A variations in GRCh37p13 or GRCh38, and will replace SNP GeneView later this year. Please visit the Help Page or YouTube for available features and send your comments and suggestions to NCBI helpdesk.

### Gene Model (mRNA alignment) information from genome sequence

<table>
<thead>
<tr>
<th>gene model (contig mRNA transcript)</th>
<th>Contig Label</th>
<th>Contig</th>
<th>mrna</th>
<th>protein</th>
<th>mrna orientation</th>
<th>transcript</th>
<th>snp count</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>NM_198056.2</td>
<td>minus strand</td>
<td>NP_932173.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>NM_001099404.1</td>
<td>minus strand</td>
<td>NP_001092874.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>NM_000335.4</td>
<td>minus strand</td>
<td>NP_000326.2</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>XM_006713284.1</td>
<td>minus strand</td>
<td>XP_006713347.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>XM_006713283.1</td>
<td>minus strand</td>
<td>XP_006713346.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>XM_006713282.1</td>
<td>minus strand</td>
<td>XP_006713345.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>XM_001160151.1</td>
<td>minus strand</td>
<td>NP_001153633.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>XM_001160150.1</td>
<td>minus strand</td>
<td>NP_001153632.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>GRCh38</td>
<td>NT_022517.19</td>
<td>NM_001099405.1</td>
<td>minus strand</td>
<td>NP_001092875.1</td>
<td>reverse</td>
<td>NT_022517.19</td>
<td>GRCh38</td>
</tr>
<tr>
<td>Region</td>
<td>Chr. position</td>
<td>mRNA pos</td>
<td>dbSNP rs# cluster id</td>
<td>Heterozygosity</td>
<td>Validation</td>
<td>MAF</td>
<td>3D Linkout</td>
</tr>
<tr>
<td>--------</td>
<td>---------------</td>
<td>----------</td>
<td>----------------------</td>
<td>----------------</td>
<td>------------</td>
<td>-----</td>
<td>-----------</td>
</tr>
<tr>
<td>38650366</td>
<td>6122</td>
<td>rs376697724</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650401</td>
<td>6162</td>
<td>rs371008670</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650430</td>
<td>6133</td>
<td>rs37569236</td>
<td>0.500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650522</td>
<td>6041</td>
<td>rs367770922</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650528</td>
<td>6035</td>
<td>rs13324203</td>
<td>0.102</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650530</td>
<td>6033</td>
<td>rs62241166</td>
<td>0.500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650564</td>
<td>5999</td>
<td>rs372582841</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650570</td>
<td>5993</td>
<td>rs375254452</td>
<td>N.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38650476</td>
<td>5087</td>
<td>rs200594132</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Validation status description**

- Validated by multiple, independent submissions to the refSNP cluster
- Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
- Validated by submitter confirmation
- All alleles have been observed in at least two chromosomes apiece
- Genotyped by HapMap project
- SNP has been sequenced in 1000Genome project.
- Suspect SNPs: SNP suspected from paralogous region (PMID: 21030649). Added to dbSNP on 01/21/2011.
### Variation Data

<table>
<thead>
<tr>
<th>Source database</th>
<th>Variant ID</th>
<th>Location</th>
<th>Variant type</th>
<th>Gene</th>
<th>Molecular consequences</th>
<th>Worst clinical significance</th>
<th>1000G MAF</th>
<th>GO-ESP MAF</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbSNP (4,158,269)</td>
<td>rs969583</td>
<td>16,324-56,449</td>
<td>copy number variation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>dbSNP (4,158,269)</td>
<td>rs676281</td>
<td>18,324-147,182</td>
<td>copy number variation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>dbSNP (4,158,269)</td>
<td>rs676282</td>
<td>16,324-170,862</td>
<td>copy number variation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Ensembl

• joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute

• launched in 1999

• centralized resource for geneticists, molecular biologists and other researchers studying the genomes of our own species and other vertebrates and model organisms

http://www.ensembl.org
Gene: SCN5A  ENSG00000183873

Description
sodium channel, voltage-gated, type V, alpha subunit [Source:HGNC Symbol;Acc:HGNC:10593]

Synonyms
CDCD2, CMD1E, CMPD2, HB1, HB2, HBBD, HH1, ICCD, IVF, LQT3, Nav1.5, PFHB1, SS81

Location
Chromosome 3: 38,548,057-38,849,673 reverse strand.

INSDC coordinates
chromosome GRCh38:CM000665:2:38548057:38649073:1

Transcripts
This gene has 15 transcripts (splice variants)  Show transcript table

Summary

Name
SCN5A (HGNC Symbol)

CCDS
This gene is a member of the Human CCDS set: CCDS45796, CCDS46797, CCDS46798, CCDS46799, CCDS54569, CCDS54570

UniprotKB
This gene has proteins that correspond to the following Uniprot identifiers: Q14624

RefSeq
Overlapping RefSeq Gene ID 6331 matches and has similar biotype of protein_coding

LRG
LRG_282 provides a stable genomic reference framework for describing sequence variations for this gene

Ensembl version
ENS000000183873.12

GRCh37 assembly
This gene maps to 38,569,548-38,691,164 in GRCh37 coordinates.

Gene type
Known protein coding

Prediction Method
Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see article

Alternative Method
This gene corresponds to the following database identifiers:
Havana gene: OTTHUMG00000156168
UCSC Genome Browser

• on-line genome browser hosted by the University of California, Santa Cruz (UCSC)

• interactive website offering access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms

• integrated with a large collection of aligned annotations

• graphical viewer optimized to support fast interactive performance and is an open-source, web-based tool suite built on top of a MySQL database for rapid visualization, examination, and querying of the data at many levels

http://genome.ucsc.edu/
GeneCards

- searchable, integrated database of human genes
- provides comprehensive, updated, and user-friendly information
- all known and predicted human genes
- extracts and integrates gene-related data:
  - Genomic
  - Transcriptomic
  - Proteomic
  - Genetic
  - Clinical
  - functional information
- Automatically mined from >100 carefully selected web sources
- Allowing one-stop access to a very broad information base

http://www.genecards.org/
SCN5A Gene
protein-coding  GIFs: 67
GCID: GC03M038589

Aliases
Sodium Channel, Voltage-Gated, Type V, Alpha Subunit
Type V, Alpha Subunit
CMD1E
Sodium Channel Protein Cardiac
Muscle Subunit Alpha
Voltage-Gated Sodium Channel Subunit Alpha Nav1.5
H123
CDCD2
HB2
LQT3

Sodium Channel, Voltage-Gated, Type V, Alpha (Long QT Syndrome 3)
CMPD2
HB2
HBDD2
ICCD2
IFV2
Nav1.5
PFHB1
If the focus is primarily SNPs....
HapMap Project: Create a genome-wide SNP map

Genotype SNPs in four populations:

- CEPH (CEU) (Europe - n = 90, trios)
- Yoruban (YRI) (Africa - n = 90, trios)
- Japanese (JPT) (Asian - n = 45)
- Chinese (HCB) (Asian - n = 45)

To produce a genome-wide map of common variation

Common Variant/Common Disease
Low density - genome-wide
Phase I - 1M SNPs
Density ~ 1 SNP/kb

Phase II - 4M SNPs

High density - candidate gene

643 genes - 15 Mbp
92,300 SNPs - 1 SNP/166 bp

322 genes - 7 Mbp
37,450 SNPs - 1 SNP/186 bp
SNP Discovery: dbSNP database

dbSNP
-NCBI SNP database
SNP data submitted to dbSNP: Clustering

**dbSNP processing of SNPs**

SNPs submitted by research community (submitted SNPs = ss#) → Unique mapping to a genome location (reference SNP = rs#) → Validated or Unvalidated

**Validation status description**
- Validated by multiple, independent submissions to the rsSNP cluster
- Validated by frequency or genotype data: minor alleles observed in at least two chromosomes.
- Validated by submitter confirmation
- All alleles have been observed in at least two chromosomes apiece
- Genotyped by HapMap project
- SNP has been sequenced in 1000Genome project
HapMap Discovery Increased SNP Density and Validated SNPs

HapMap SNP Discovery

- 14+ million rs SNPs
- 6.5 million validated rs SNPs
rs #’s are THE nomenclature for SNPs

Table 1: Association between SNPs in the chromosome 20 locus and AGA in the German sample

<table>
<thead>
<tr>
<th>SNP (position)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GWAS</td>
</tr>
<tr>
<td>Replication</td>
</tr>
<tr>
<td>Combined§</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample</th>
<th>Cases</th>
<th>Controls</th>
<th>MAFa</th>
<th>Genotypesb</th>
<th>P</th>
<th>OR (95% CI)f</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cases</td>
<td>Controls</td>
<td>Cases</td>
<td>Controls</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6137444</td>
<td>296</td>
<td>347</td>
<td>0.264(C)</td>
<td>0.383(C)</td>
<td>14/128/154</td>
<td>49/168/130</td>
</tr>
<tr>
<td>(21,733,639 bp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21/135/163</td>
<td>45/99/90</td>
</tr>
<tr>
<td>rs2180439</td>
<td>296</td>
<td>347</td>
<td>0.292(C)</td>
<td>0.429(C)</td>
<td>21/131/144</td>
<td>66/166/115</td>
</tr>
<tr>
<td>(21,801,100 bp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23/147/149</td>
<td>62/103/69</td>
</tr>
<tr>
<td>rs1998076</td>
<td>296</td>
<td>347</td>
<td>0.282(A)</td>
<td>0.427(A)</td>
<td>20/120/144</td>
<td>65/163/115</td>
</tr>
<tr>
<td>(21,828,045 bp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23/146/150</td>
<td>61/102/71</td>
</tr>
<tr>
<td>rs201571</td>
<td>296</td>
<td>347</td>
<td>0.292(A)</td>
<td>0.448(A)</td>
<td>43/267/295</td>
<td>125/259/183</td>
</tr>
<tr>
<td>(21,961,514 bp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17/137/142</td>
<td>61/163/123</td>
</tr>
<tr>
<td>rs6113491</td>
<td>296</td>
<td>347</td>
<td>0.359(C)</td>
<td>0.483(C)</td>
<td>29/154/112</td>
<td>88/159/100</td>
</tr>
<tr>
<td>(22,005,415 bp)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38/156/125</td>
<td>77/105/52</td>
</tr>
</tbody>
</table>

Increasing SNP Density: HapMap ENCODE Project

**ENCODE = ENCyclopedia Of DNA Elements**
Catalog all functional elements in 1% of the genome (30 Mb)

10 Regions x 500 kb/region (Pilot Project)
David Altschuler (Broad), Richard Gibbs (Baylor)
16 CEU, 16 YRI, 8 HCB, 8 JPT

Comprehensive PCR based resequencing across these regions

<table>
<thead>
<tr>
<th>Region name</th>
<th>Chromosome band</th>
<th>Genomic interval (NCBI)</th>
<th>Available SNPs</th>
<th>dbSNP</th>
<th>New SNPs</th>
<th>Gem</th>
<th>no rs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENr112</td>
<td>2p16.3</td>
<td>Chr2:51633239..52133238</td>
<td>1,624</td>
<td>1,720</td>
<td></td>
<td></td>
<td>1,084</td>
</tr>
<tr>
<td>ENr131</td>
<td>2q37.1</td>
<td>Chr2:234778639..235278638</td>
<td>1,787</td>
<td>1,233</td>
<td></td>
<td></td>
<td>1,179</td>
</tr>
<tr>
<td>ENr113</td>
<td>4q26</td>
<td>Chr4:118705475..119265474</td>
<td>1,516</td>
<td>1,819</td>
<td></td>
<td></td>
<td>1,017</td>
</tr>
<tr>
<td>ENm010</td>
<td>7p15.2</td>
<td>Chr7:26699793..27199792</td>
<td>1,274</td>
<td>1,857</td>
<td></td>
<td></td>
<td>757</td>
</tr>
<tr>
<td>ENm013</td>
<td>7q21.13</td>
<td>Chr7:89395718..89895717</td>
<td>1,545</td>
<td>1,713</td>
<td></td>
<td></td>
<td>927</td>
</tr>
<tr>
<td>ENm014</td>
<td>7q31.33</td>
<td>Chr7:126135436..126632577</td>
<td>1,354</td>
<td>1,562</td>
<td></td>
<td></td>
<td>963</td>
</tr>
<tr>
<td>ENr321</td>
<td>8q24.11</td>
<td>Chr8:118709628..119269627</td>
<td>1,468</td>
<td>1,682</td>
<td></td>
<td></td>
<td>936</td>
</tr>
<tr>
<td>ENr232</td>
<td>9q34.11</td>
<td>Chr9:127081347..127561346</td>
<td>1,494</td>
<td>1,646</td>
<td></td>
<td></td>
<td>694</td>
</tr>
<tr>
<td>ENr123</td>
<td>12q12</td>
<td>Chr12:238626477..39126476</td>
<td>1,904</td>
<td>1,551</td>
<td></td>
<td></td>
<td>859</td>
</tr>
<tr>
<td>ENr213</td>
<td>18q12.1</td>
<td>Chr18:23717221..24217220</td>
<td>1,391</td>
<td>1,485</td>
<td></td>
<td></td>
<td>809</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>15,357</td>
<td>16,248</td>
<td></td>
<td></td>
<td>9,205</td>
</tr>
</tbody>
</table>

15,357 dbSNP
16,248 New SNPs
50% of SNPs in dbSNP
5 Mb/31,500 SNPs = 1/160 bp

Population descriptors:
- **CEU**: CEPH (Utah residents with ancestry from northern and western Europe)
- **HCB**: Han Chinese in Beijing, China
- **JPT**: Japanese in Tokyo, Japan
- **YRI**: Yoruba in Ibadan, Nigeria
Goal:
Comprehensively identify all common sequence variation in candidate genes

Initial biological focus:
Candidate environmental response genes involved in DNA repair, cell cycle, apoptosis, metabolism, cell signaling, and oxidative stress.

Approach:
Direct resequencing of genes

Samples:
PDR-90 ethnically diverse individuals representative of U.S. population (397 genes)
EGP95-95 samples from four ethnic groups (227 genes)
(24 HapMap Asians, 22 HapMap Europeans, 12 HapMap Yorubans, 15 African Americans, 22 Hispanics)

Website:
egp.gs.washington.edu
Goal: Comprehensively identify all common sequence variation in candidate genes

Initial biological focus: Candidate environmental response genes involved in lipid metabolism, inflammation, and blood pressure regulation.

Approach: Direct resequencing of genes

Samples:
P1: 23 CEPHs and 24 African-American (overlaps with Perlegen)  
P2: 23 CEPHs and 24 Yorubans (overlaps with HapMap)

Website: pga.gs.washington.edu
643 genes sequenced (NIEHS SNPs)  
15 Mb scanned  
> 92,000 genotyped SNPs identified  
> 8 million genotypes deposited in dbSNP
Approximately 10 million common SNPs exist in the human genome (1/300 bp).

Random SNP discovery processes generate many SNPs (HapMap).

Random approaches to SNP discovery have reached limits of discovery and validation (1/600 bp; 50% SNP validation).

Most validated SNPs (6+ million) have been genotyped by the HapMap (3 pops).

Resequencing approaches continue to catalog important variants (rare and common not captured by the HapMap).
1000 genomes project: motivation

- GWAS shows that systematic association studies can be used to map disease genes
- The first generation of GWAS was well powered only for SNPs with > 5% MAF
- Next generation sequencing now makes it possible to create a complete catalogue of human polymorphism for SNPs and CNVs
Exploring the full range of genetic variants

<table>
<thead>
<tr>
<th>Allele Frequency</th>
<th>High-frequency polymorphisms</th>
<th>Eg: many now known</th>
<th>HapMap</th>
<th>First generation arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rarer Alleles, Stronger Effects

<table>
<thead>
<tr>
<th>Rare Mutations</th>
<th>Eg: most mendelian</th>
<th>Direct sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MC4R, ABCA1</td>
<td>Array-based detection (CNV)</td>
</tr>
<tr>
<td>lq21.1 in SCZ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Exploring the full range of genetic variants

<table>
<thead>
<tr>
<th>Allele Frequency</th>
<th>High-frequency polymorphisms</th>
<th>Eg: many now known</th>
<th>HapMap First generation arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rarer Alleles, Stronger Effects</td>
<td>Lower-frequency polymorphisms</td>
<td>Eg: CFTR delta 508, PCSK9 C679X</td>
<td>1000 Genomes Project New arrays, imputation</td>
</tr>
<tr>
<td>Rare Mutations</td>
<td>Eg: most mendelian</td>
<td>MC4R, ABCA1</td>
<td>Direct sequencing Array-based detection (CNV)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1q21.1 in SCZ</td>
<td></td>
</tr>
</tbody>
</table>
Exploring the full range of genetic variants

Allele Frequency

50%  | High-frequency polymorphisms  | Eg: many now known  | HapMap  
|     |                              |                     | First generation arrays |

Rarer Alleles, Stronger Effects

5%   | Lower-frequency polymorphisms  | Eg: CFTR delta 508, PCSK9 C679X | 1000 Genomes Project  
|     |                              |                     | New arrays, imputation |

0.5% | Rare Mutations  | Eg: most mendelian  
|     |                 | MC4R, ABCA1  
|     |                 | 1q21.1 in SCZ  |

0.05% | Direct sequencing  
|      | Array-based detection (CNV)  |
1000 Genomes Project

Random Coverage
0.2 to 0.4 X depth

Produce a catalog of variants across the genome in multiple populations with allele frequencies > 1%
Where to find SNPs and Linkage Disequilibrium Data

For your gene or region of interest, search

**Genome Variation Server**

- HapMap
  www.hapmap.org
- NIEHS SNPs
  egp.gs.washington.edu
- SeattleSNPs PGA
  pga.gs.washington.edu
Visualizing Pair-wise LD

SeattleSNPs
Variation Discovery Resource

Welcome to SeattleSNPs

SeattleSNPs is funded as part of the National Heart Lung and Blood Institute’s (NHLBI) Programs for Genomic Applications (PGA). The SeattleSNPs PGA is focused on identifying, genotyping, and modeling the associations between single nucleotide polymorphisms (SNPs) in candidate genes and pathways that underlie inflammatory responses in humans.

Investigator Opportunities

SeattleSNPs offers investigators several opportunities to make use of the project’s resources:

Nominate Genes for Resequencing

As part of its mission, SeattleSNPs is soliciting requests from individual investigators for candidate genes to be resequenced for SNP discovery.

Traveling Workshops

SeattleSNPs is now accepting applications from potential host sites for One- and Two-Day Traveling Workshops.

Genotyping

SeattleSNPs is providing genotyping support for research related to heart, lung, blood, and sleep.

Latest Updates

- [New Workshop Presentation](http://example.com) added on August 17, 2009
- [NHP added to Finished Genes Page](http://example.com) Jul 19, 2009
- [PGA Case Western Reserve University](http://example.com) added on April 10, 2008
- [GPR added to Finished Genes Page](http://example.com) Feb 14, 2008
- [GPR added to Finished Genes Page](http://example.com) Dec 4, 2007
- [PRR added to Finished Genes Page](http://example.com) Dec 4, 2007
- [KPR added to Finished Genes Page](http://example.com) Nov 14, 2007
- [PMP added to Finished Genes Page](http://example.com) Nov 14, 2007
- [KPR added to Finished Genes Page](http://example.com) Oct 12, 2007
- [CSN Clinical Cardiovascular Genomics Meeting](http://example.com) added on October 10, 2007
- [GEM added to Finished Genes Page](http://example.com) Jul 25, 2007
- [GEM added to Finished Genes Page](http://example.com) Jul 22, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 31, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 20, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 19, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 8, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 5, 2007
- [GEM added to Finished Genes Page](http://example.com) Jun 4, 2007
Catalog of SNP effects

SNPedia

SNPedia is a wiki investigating human genetics. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. It is used by help explain your DNA.

Help!

- look at the example rs1234
- learn more about SNPs
- browse
  - genes
  - genomes
  - genosets
  - genotypes
  - medicines
  - medical conditions

Popular

- rs63676 in the oxytocin receptor influences social behavior and personality
- rs1816739
- rs7412 and rs423350 can raise the risk of Alzheimer's disease by more than 10x
- rs6152 can influence baldness
- rs333 resistance to HIV
- rs1800497 in a dopamine receptor may influence the sense of pleasure
- rs18005007 determines red hair and sensitivity to anesthetics
SNP-related Websites

- SeattleSNPs (pga.gs.washington.edu)
- NIEHS SNPs (egp.gs.washington.edu)
- Genome Variation Server (http://gvs.gs.washington.edu/GVS/)
- HapMap (www.hapmap.org)
- SNPedia (www.snpedia.com)
Assignment

• Search your favorite gene in the databases discussed today

• If you do not have a favorite gene, try one of mine:
  • SCN5A
  • RYR1
  • CETP
  • PCSK9
  • FTO
  • CDKN2B
  • PTPN22
Questions???