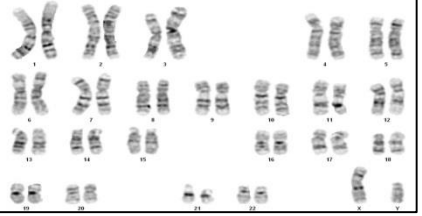
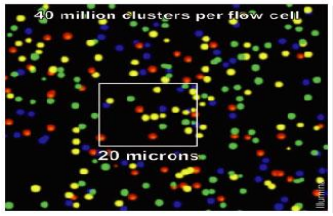
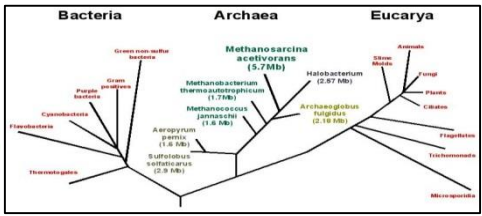
生物信息学：导论与方法
Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University

https://www.coursera.org/course/pkubioinfo





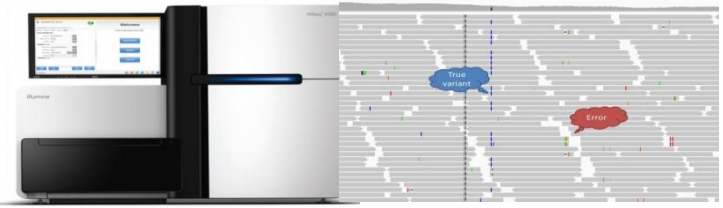


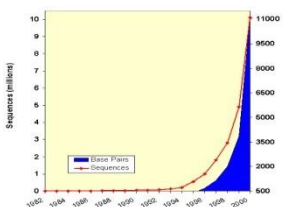
Week 6: Functional prediction of genetic variations

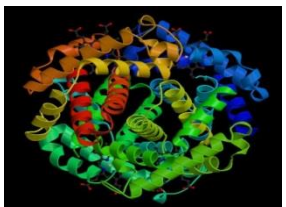
北京大学生物信息学中心 魏丽萍

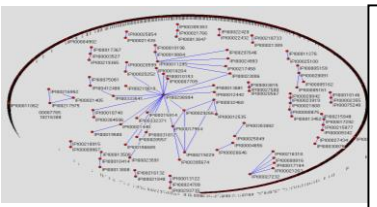
Liping Wei, Ph.D.

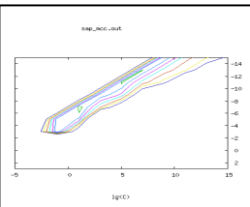
Center for Bioinformatics, Peking University

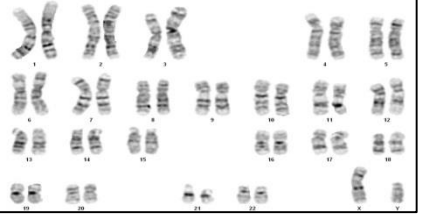


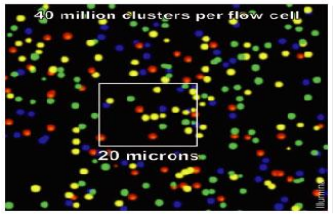


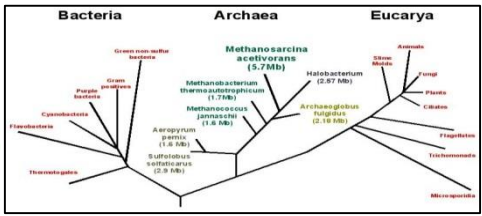










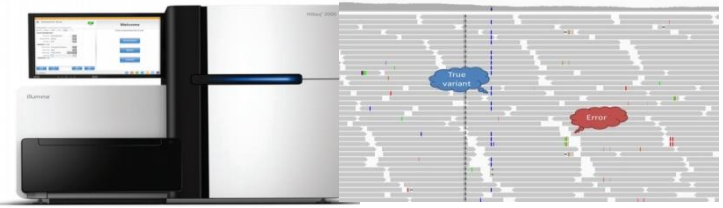


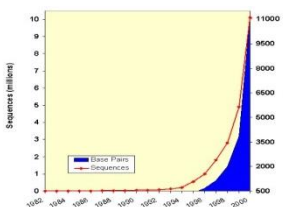
Unit 2: Databases of genetic variations

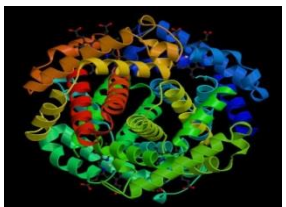
北京大学生物信息学中心 魏丽萍

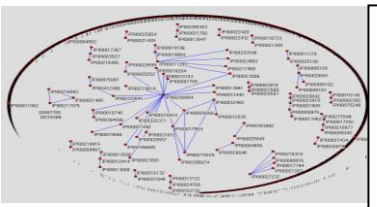
Liping Wei, Ph.D.

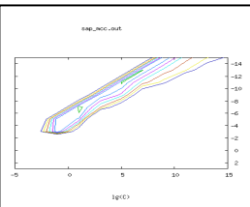
Center for Bioinformatics, Peking University

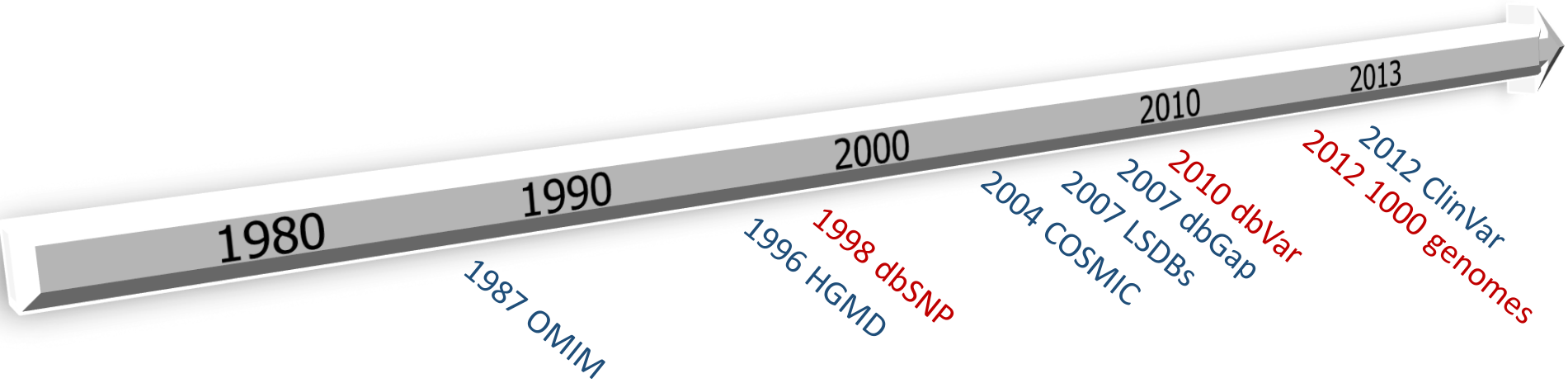











1980
1990
2000
2010
2013
1987 OMIM
1996 HGMD
1998 dbSNP
2004 COSMIC
2007 LSDBs
2007 dbGap
2010 dbVar
2012 1000 genomes
2012 ClinVar

dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)

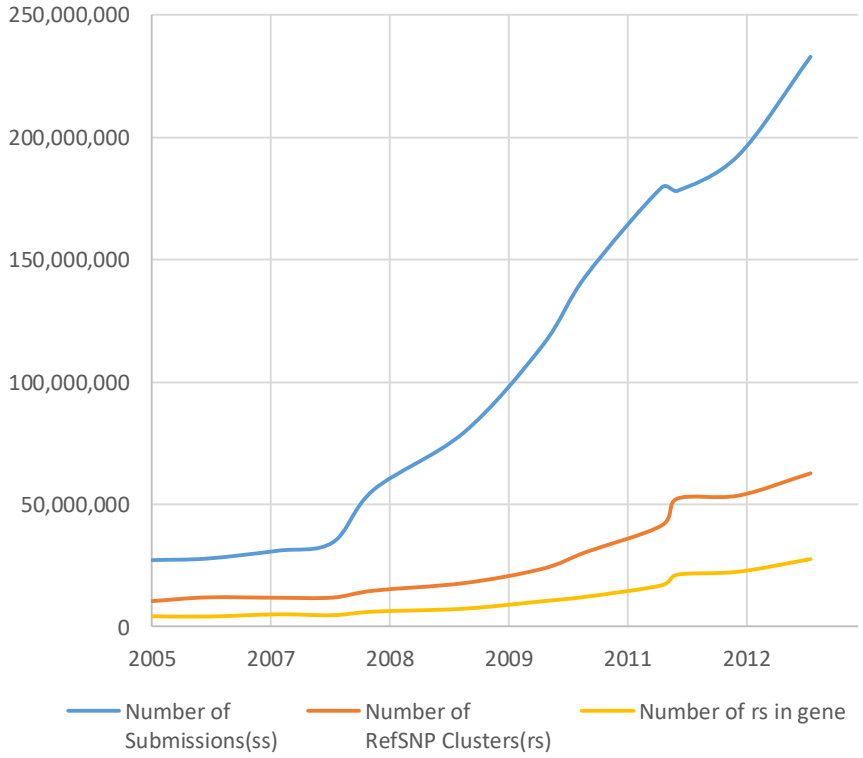
dbSNP build 138 contains genetic variations from131 species

- SNPs
- Indels
- multinucleotide polymorphisms
- microsatellite markers
- short tandem repeats
- heterozygous sequences

In *Homo sapiens:*	
Number of Submissions (ss)	232,952,851
Number of RefSNP clusters (rs)	62,676,337
Validated rs	44,278,189
Number of rs in gene	27,608,151
Number of ss with genotype	73,909,251
Number of ss with frequency	35,997,830

dbSNP – Data increase

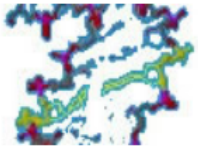
From dbSNP build 125 in 2005 to build 138 in 2013, for *Homo sapiens*



ftp://ftp.ncbi.nih.gov/pub/factsheets/Factsheet_SNP.pdf

NCBI

dbSNP
Short Genetic Variations



PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

- GENERAL
- HUMAN VARIATION
  - Search, Annotate, Submit
  - Annotate and Submit Batch Data with Clinical Impact Attributes for Filtering Variation
- SNP SUBMISSION
- DOCUMENTATION
- SEARCH
- RELATED SITES

Reference SNP(refSNP) Cluster Report: rs1800730 ** With patho

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	89/138
Map to Genome Build:	37.5
Validation Status:	

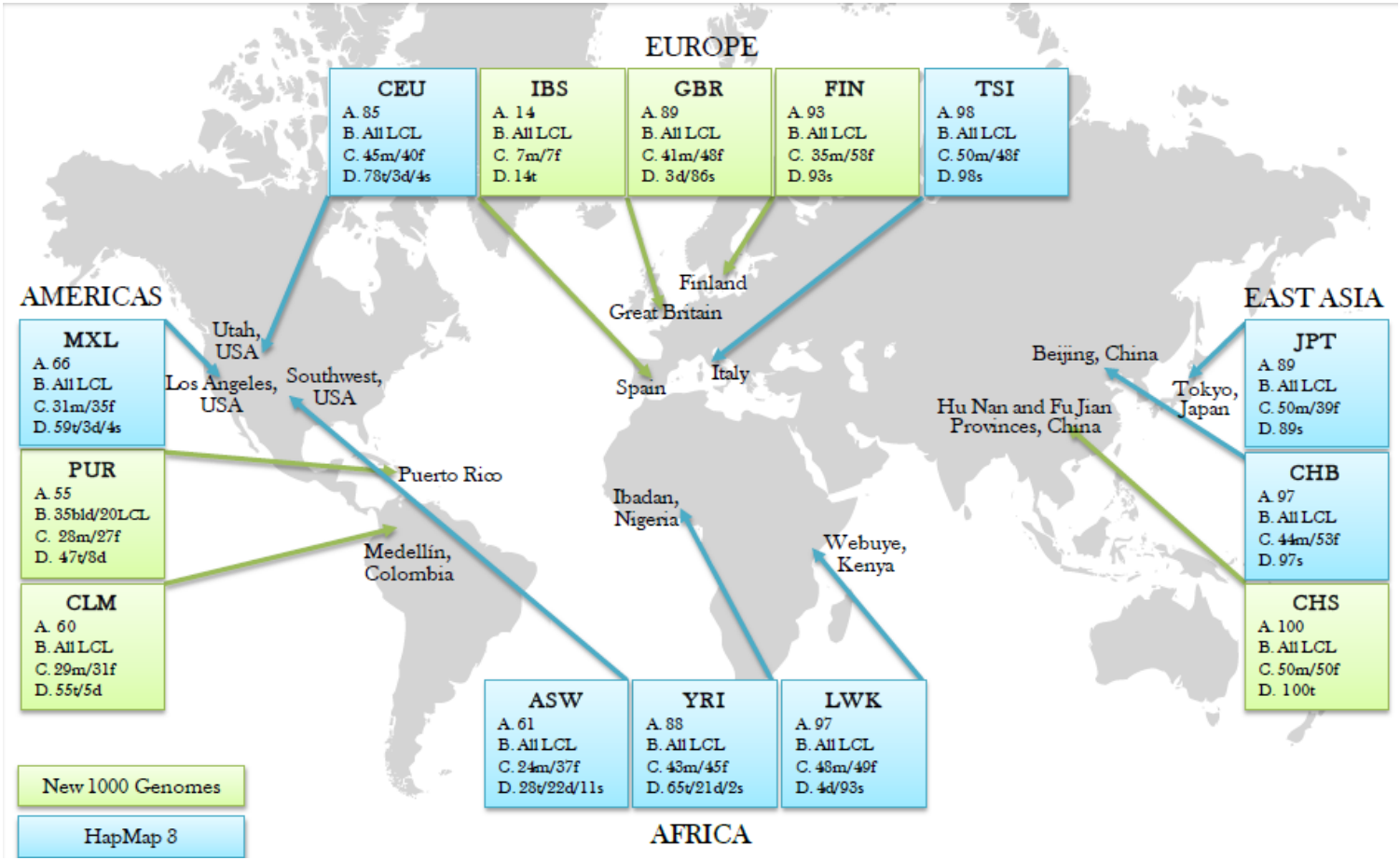
Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	A/T
Allele Origin:	A:germline T:germline
Ancestral Allele:	A
Clinical Channel:	VarView OMIM
Clinical Significance:	With pathogenic allele [detail]
MAF/MinorAlleleCount:	T=0.007/16
MAF Source:	1000 Genomes

HGVS Names
NC_000006.11:g.26091185A>T
NG_008720.1:g.8677A>T
NM_000410.3:c.193A>T
NM_139003.2:c.193A>T
NM_139004.2:c.193A>T
NM_139006.2:c.193A>T
NM_139007.2:c.77-357A>T
NM_139008.2:c.77-357A>T
NM_139009.2:c.124A>T
NM_139010.2:c.77-1728A>T
NM_139011.2:c.77-2162A>T
NP_000401.1:p.Ser65Cys
NP_620572.1:p.Ser65Cys
NP_620573.1:p.Ser65Cys
NP_620575.1:p.Ser65Cys
NP_620578.1:p.Ser42Cys
NT_007592.15:g.26031185A>T

SNP Details are organized in the following sections:

GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

1000 Genomes (http://www.1000genomes.org/)



1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65.

1000 Genomes

- Illumina
- SOLID
- 454

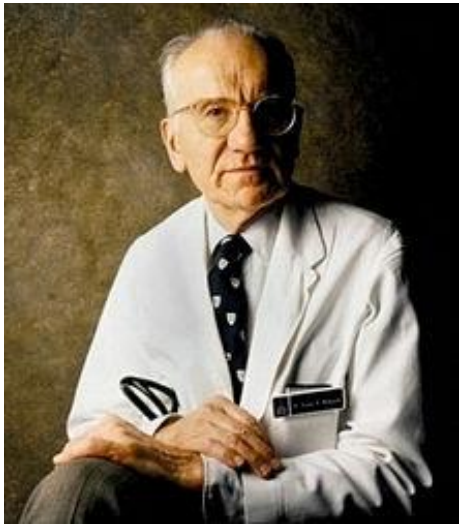
Phase I	strategy	Coverage	Sample number
Whole genome	Low-coverage whole-genome sequencing	2-6X	1,092
Whole exome	Deep sequencing of whole exomes	50-100X	1,039

	Autosomes	Chromosome X
Samples	1,092	1,092
Total raw bases (Gb)	19,049	804
Mean mapped depth (×)	5.1	3.9
SNPs		
No. sites overall	36.7 M	1.3 M
Novelty rate†	58%	77%
No. synonymous/non-synonymous/nonsense	NA	4.7/6.5/0.097 K
Average no. SNPs per sample	3.60 M	105 K
Indels		
No. sites overall	1.38 M	59 K
Novelty rate†	62%	73%
No. inframe/frameshift	NA	19/14
Average no. indels per sample	344 K	13 K
Genotyped large deletions		
No. sites overall	13.8 K	432
Novelty rate†	54%	54%
Average no. variants per sample	717	26

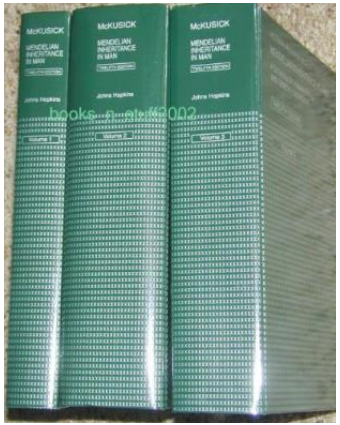
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	
1	10583	rs58108140		G	A	100	PASS
1	10611	rs189107123		C	G	100	PASS
1	13302	rs180734498		C	T	100	PASS
1	13327	rs144762171		G	C	100	PASS
1	13957	rs201747181		TC	T	28	PASS
1	13980	rs151276478		T	C	100	PASS
1	30923	rs140337953		G	T	100	PASS
1	46402	rs199681827		C	CTGT	31	PASS
1	47190	rs200430748		G	GA	192	PASS
1	51476	rs187298206		T	C	100	PASS
1	51479	rs116400033		T	A	100	PASS
1	51914	rs190452223		T	G	100	PASS
1	51935	rs181754315		C	T	100	PASS
1	51954	rs185832753		G	C	100	PASS
1	52058	rs62637813		G	C	100	PASS
1	52144	rs190291950		T	A	100	PASS

OMIM (Online Mendelian Inheritance in Man)

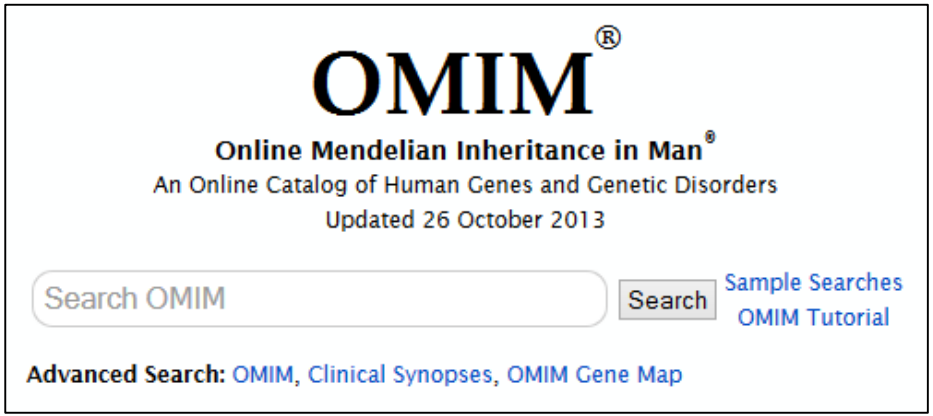
http://www.omim.org/



Victor A. McKusick



MIM

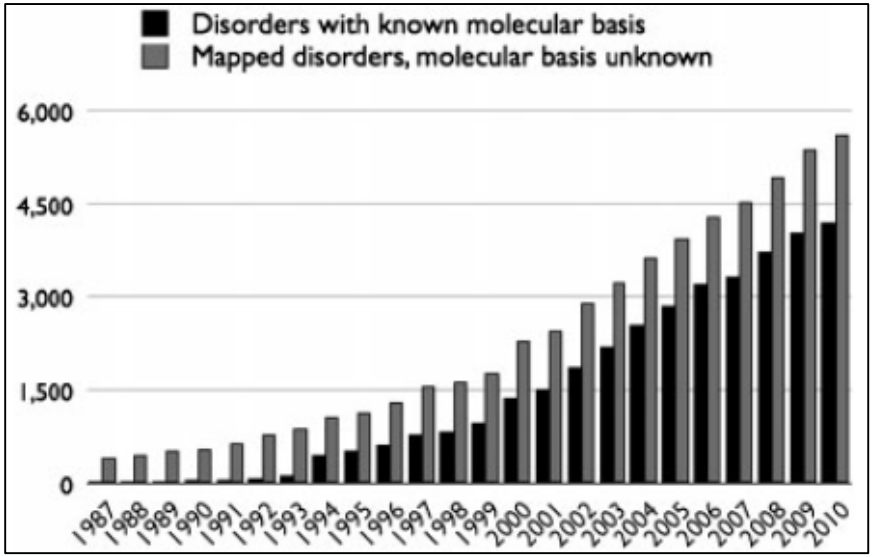


OMIM

http://www.amazon.com/Mendelian-Inheritance-Man-Catalog-Disorders/dp/0801857422

http://en.wikipedia.org/wiki/Victor_A._McKusick

OMIM Statistics



Number of Entries in OMIM (Updated 26 October 2013) :

Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
* Gene description	13,662	666	48	35	14,411
+ Gene and phenotype, combined	111	3	0	2	116
# Phenotype description, molecular basis known	3,625	280	4	28	3,937
% Phenotype description or locus, molecular basis unknown	1,591	131	5	0	1,727
Other, mainly phenotypes with suspected mendelian basis	1,755	118	2	0	1,875
Totals	20,744	1,198	59	65	22,066

Amberger J et al. A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). Hum Mutat. 2011 May;32(5):564-7.

OMIM

breast cancer Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map **Toggle:** search terms highlighted
Search History: View, Clear

#114480

BREAST CANCER

Alternative titles; symbols

BREAST CANCER, FAMILIAL

Other entities represented in this entry:

BREAST CANCER, FAMILIAL MALE, INCLUDED

Phenotype Gene Relationships

Clinical Synopsis

TEXT

A number sign (#) is used with this entry because of evidence that mutation at more tha
can be involved in different families or even in the same case. These loci include BRCA1
17q, BRCA2 (600185) on 13q12, BRCATA (600048) on 11q, BRCA3 (605365) on 13q21,
(602631) on 11p15.5, the TP53 gene (191170) on 17p, and the RB1CC1 gene (60683
Mutations in the androgen receptor gene (AR; 313700) on the X chromosome have be
cases of male breast cancer (313700.0016). Mutation in the RAD51 gene (179617) wa
patients with familial breast cancer (179617.0001). Breast cancer susceptibility alleles
reported in the CHEK2 gene (see 604373.0001 and 604373.0012) and in the BARD
601593.0001).

BRCA1 Search Sort by: Relevance Date updated

Advanced Search: OMIM, Clinical Synopses, OMIM Gene Map **Toggle:** search terms highlighted , changes highlighted
Search History: View, Clear

*113705

BREAST CANCER 1 GENE; BRCA1

HGNC Approved Gene Symbol: *BRCA1*

Cytogenetic location: *17q21.31* ***Genomic coordinates (GRCh37):*** *17:41,196,311 - 41,277,499* (from NCBI)

Gene Phenotype Relationships

Location	Phenotype	Phenotype MIM number
17q21.31	{Breast-ovarian cancer, familial, 1}	604370
	{Pancreatic cancer, susceptibility to, 4}	614320

- Table of Contents - *113705

External Links:

- Genome
- DNA
- Protein
- Gene Info
- Clinical Resources
- Variation
- Animal Models
- Cellular Pathways

TEXT

Description

BRCA1 plays critical roles in DNA repair, cell cycle checkpoint control, and maintenance of genomic stability. BRCA1 forms several distinct complexes through association with different adaptor proteins, and each complex forms in a mutually exclusive manner (Wang et al., 2009).

Cloning

Miki et al. (1994) identified cDNA sequences corresponding to the BRCA1 gene by positional cloning of the region on 17q21 implicated in familial breast-ovarian cancer syndrome (604370). The deduced 1,863-residue protein with zinc-finger domains near the N terminus. A 7.8-kb mRNA transcript was identified in testes, thymus, breast and ovary. There appeared to be a complex pattern of alternative splicing.

Bennett et al. (1995) found that the mouse Brca1 gene shares 75% identity of the coding region with

Human Gene Mutation Database (HGMD)

www.hgmd.cf.ac.uk/

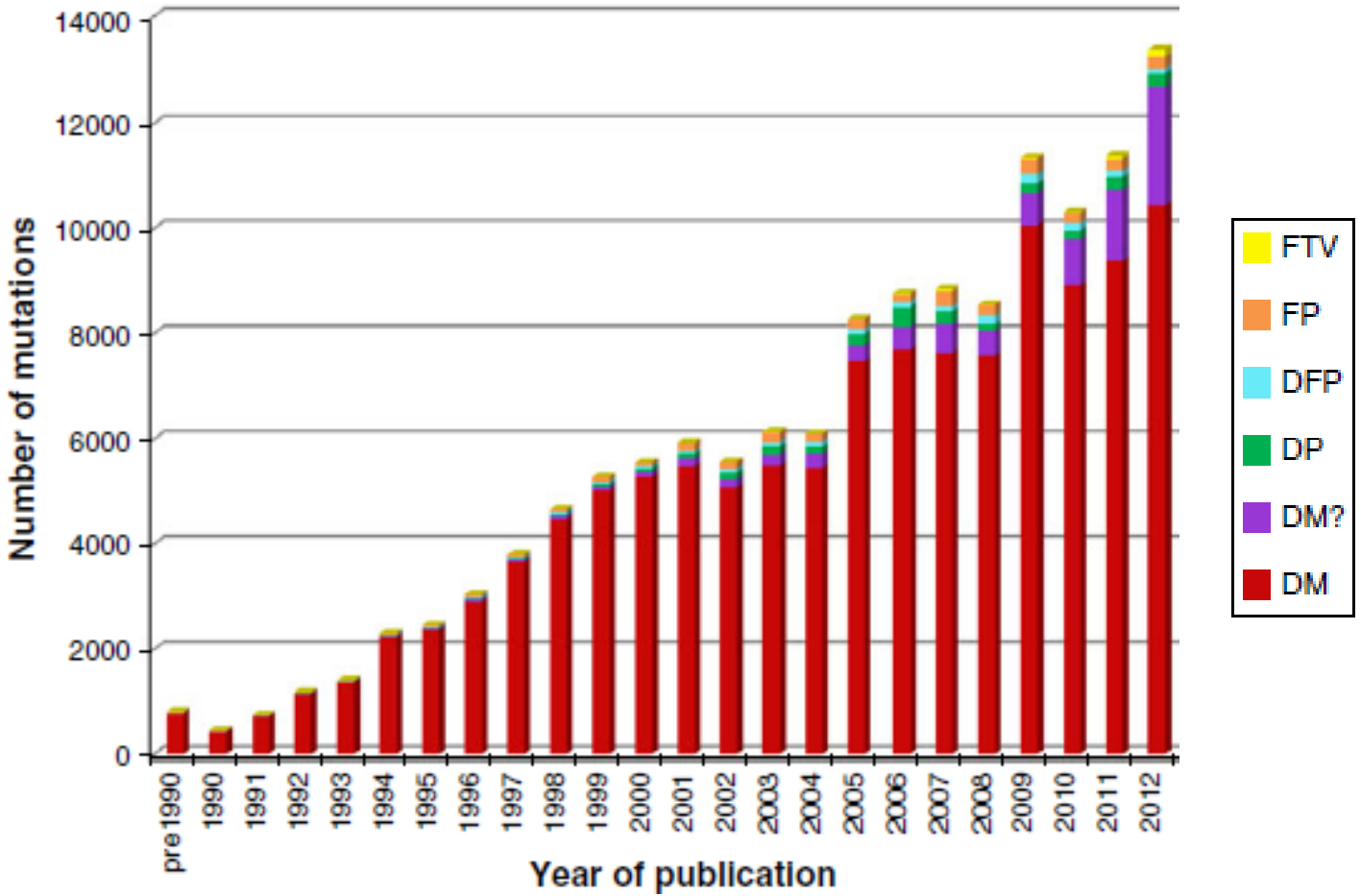
- a comprehensive collection of gene mutations that underlie, or are associated with, human genetic diseases, manually curated from literature.

http://www.hgmd.cf.ac.uk/ac/index.php

David N. Cooper

Michael Krawczak

HGMD



Stenson PD et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2013 Sep 28.

HGMD (February, 2013)

Mutation type	Total numbers of mutations		
	HGMD Professional	With chromosomal coordinates	Publicly available
Missense substitutions	62,368	61,845	44,933
Nonsense substitutions	15,781	15,574	11,306
Splicing substitutions	13,030	12,538	9,467
Regulatory substitutions	2,751	2,713	1,753
Micro-deletions $\leq$ 20 bp	21,681	21,134	15,796
Micro-insertions $\leq$20 bp	8,994	8,721	6,494
Micro-indels $\leq$20 bp	2,083	2,004	1,459
Gross deletions >20 bp	10,267	0	6,156
Gross insertions/ duplications >20 bp	2,376	0	1,253
Complex rearrangements	1,409	0	946
Repeat variations	421	0	305
Totals	141,161	124,529	99,868

Stenson PD et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2013 Sep 28.

NM_007294.3

Gene symbol: BRCA1

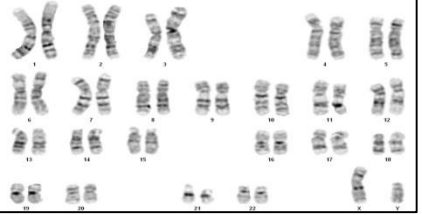
Extended cDNA not available

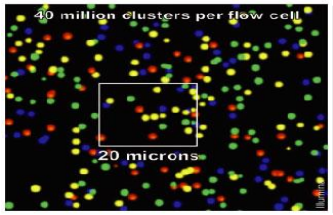
Database: Missense/nonsense - Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet. There are currently **422** mutations available in this category.

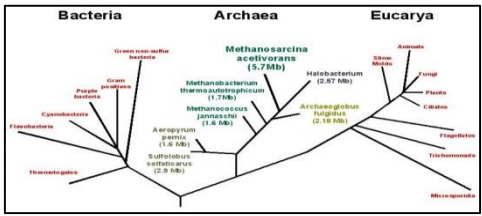
Missense/nonsense	Splicing	Regulatory	Small deletions	Small insertions	Small indels	Gross deletions	Gross insertions	Complex	Repeats
522 mutations in HGMD professional 2013.1	**150** mutations in HGMD professional 2013.1	**11** mutations in HGMD professional 2013.1	**510** mutations in HGMD professional 2013.1	**169** mutations in HGMD professional 2013.1	**29** mutations in HGMD professional 2013.1	**181** mutations in HGMD professional 2013.1	**37** mutations in HGMD professional 2013.1	**21** mutations in HGMD professional 2013.1	**1** mutation in HGMD professional 2013.1

Further options available in HGMD professional 2013.1

Accession Number	Codon change	Amino acid change	Codon number	Genomic coordinates & HGVS nomenclature	Phenotype	Reference	Comments
CM021503	aATG-GTG	Met-Val	1	BIOBASE Feature available to subscribers	Breast and/or ovarian cancer	Meindl (2002) Int J Cancer **97,** 472 Additional phenotype report available to subscribers	aka 120 A>G.
CM041678	ATG-ACG	Met-Thr	1	BIOBASE Feature available to subscribers	Breast and/or ovarian cancer ?	Abkevich (2004) J Med Genet **41,** 492	
CM014520	ATG-AGG	Met-Arg	1	BIOBASE Feature available to subscribers	Ovarian cancer	Sekine (2001) Clin Cancer Res **7,** 3144 Additional report available to subscribers Functional characterisation report available to subscribers	
CM960163	ATGg-ATT	Met-Ile	1	BIOBASE Feature available to subscribers	Breast cancer	Couch (1996) Hum Mutat **8,** 8 Additional report available to subscribers Additional phenotype report available to subscribers	
CM940170	GTA-GCA	Val-Ala	11	BIOBASE Feature available to subscribers	Breast cancer	Castilla (1994) Nat Genet **8,** 387	
CM031646	aCAA-TAA	Gln-Term	12	BIOBASE Feature available to subscribers	Breast cancer	Adem (2003) Cancer **97,** 1	
CM041679	ATT-ACT	Ile-Thr	15	BIOBASE Feature available to subscribers	Breast and/or ovarian cancer ?	Abkevich (2004) J Med Genet **41,** 492	
CM012906	ATG-AAG	Met-Lys	18	BIOBASE Feature available to subscribers	Ovarian cancer	Machackova (2001) Hum Mutat **18,** 545	





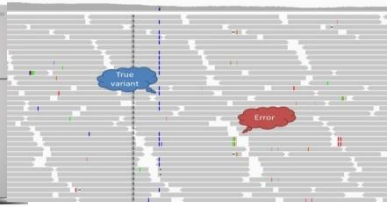


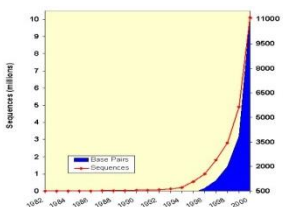
Unit 3: Conservation-based and Rule-based methods: SIFT & PolyPhen

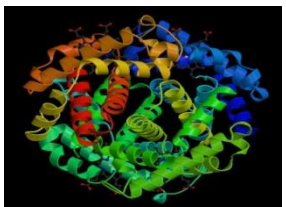
北京大学生物信息学中心 魏丽萍

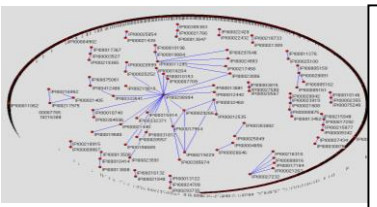
Liping Wei, Ph.D.

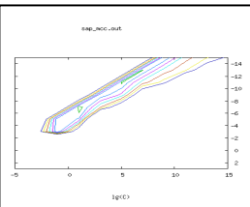
Center for Bioinformatics, Peking University











生物信息学：导论与方法
Bioinformatics: Introduction and Methods

Ge Gao 高歌 & Liping Wei 魏丽萍
Center for Bioinformatics, Peking University

https://www.coursera.org/course/pkubioinfo