

蔡涛

caitao@nibs.ac.cn

我眼中的生物信息学

Bioinformatics = Data + Algorithm

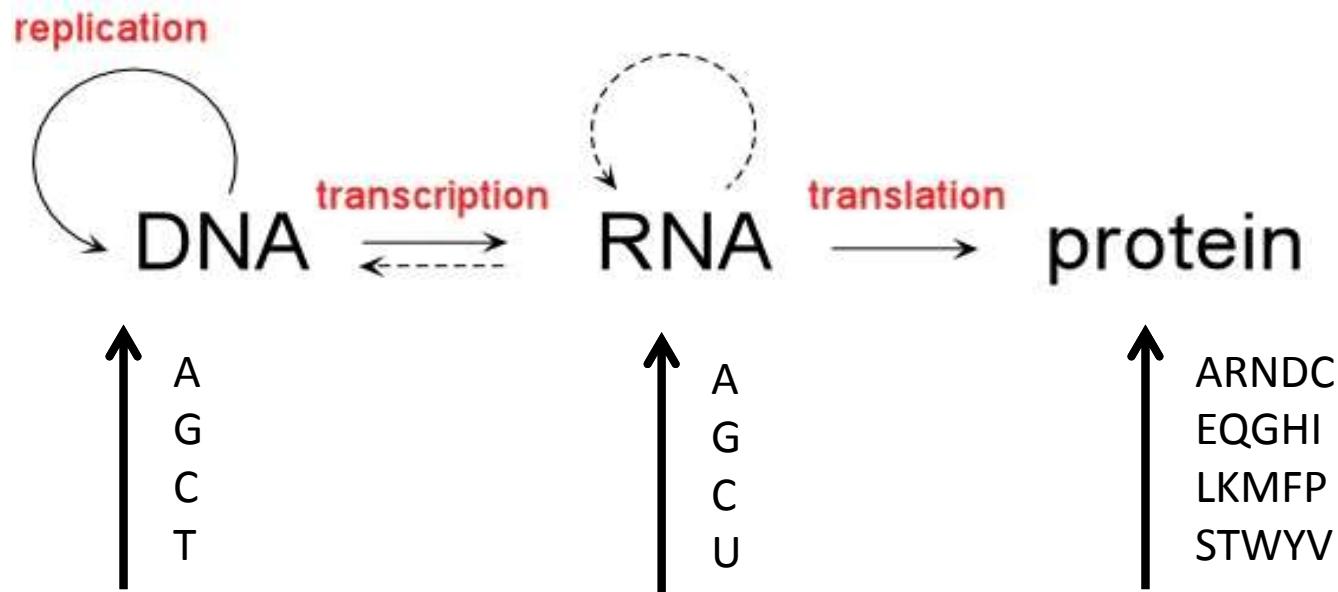


Bioinformatics

Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

--NIH Bioinformatics Definition Committee

The Central dogma



Primary database

- NCBI GenBank /USA
- EMBL-EBI resource /EU
- DDBJ /Japan

NCBI Entrez interface

The screenshot shows the NCBI Entrez interface for the gene INS (insulin) in *Homo sapiens* (human). The search bar at the top has "Gene" selected. The main content area displays the gene summary, including its official symbol (INS), full name (insulin), primary source (HGNC:HGNC:6081), and various links to related databases like Ensembl, HPRD, MIM, and Vega. It also shows the gene type (protein coding), RefSeq status (REVIEWED), organism (*Homo sapiens*), lineage (Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo), and other identifiers like IDDM, ILPR, IRDN, IDDM1, IDDM2, MODY10, and several MIM numbers. A summary section describes the post-translational cleavage of proinsulin into insulin and C-peptide. On the right, a sidebar provides links to the Table of contents, Summary, Genomic context, Genomic regions, transcripts, and products, Bibliography, Phenotypes, Variation, HIV-1 interactions, Pathways from BioSystems, Interactions, General gene information (Markers, Readthrough INS-IGF2, Homology, Gene Ontology), General protein information, NCBI Reference Sequences (RefSeq), Related sequences, and Additional links.

NCBI Resources How To Sign in to NCBI

Gene Gene Advanced Search Help

Display Settings: Full Report Send to:

Hide sidebar >>

INS insulin [*Homo sapiens* (human)]

Gene ID: 3630, updated on 17-Mar-2015

Summary

Official Symbol INS provided by HGNC

Official Full Name insulin provided by HGNC

Primary source HGNC:HGNC:6081

See related Ensembl:ENSG00000254647; HPRD:01455; MIM:176730; Vega:OTTHUMG00000009558

Gene type protein coding

RefSeq status REVIEWED

Organism *Homo sapiens*

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as IDDM; ILPR; IRDN; IDDM1; IDDM2; MODY10

Summary After removal of the precursor signal peptide, proinsulin is post-translationally cleaved into three peptides: the B chain and A chain peptides, which are covalently linked via two disulfide bonds to form insulin, and C-peptide. Binding of insulin to the insulin receptor (INSR) stimulates glucose uptake. A multitude of

Table of contents

Summary

Genomic context

Genomic regions, transcripts, and products

Bibliography

Phenotypes

Variation

HIV-1 interactions

Pathways from BioSystems

Interactions

General gene information

Markers, Readthrough INS-IGF2, Homology, Gene Ontology

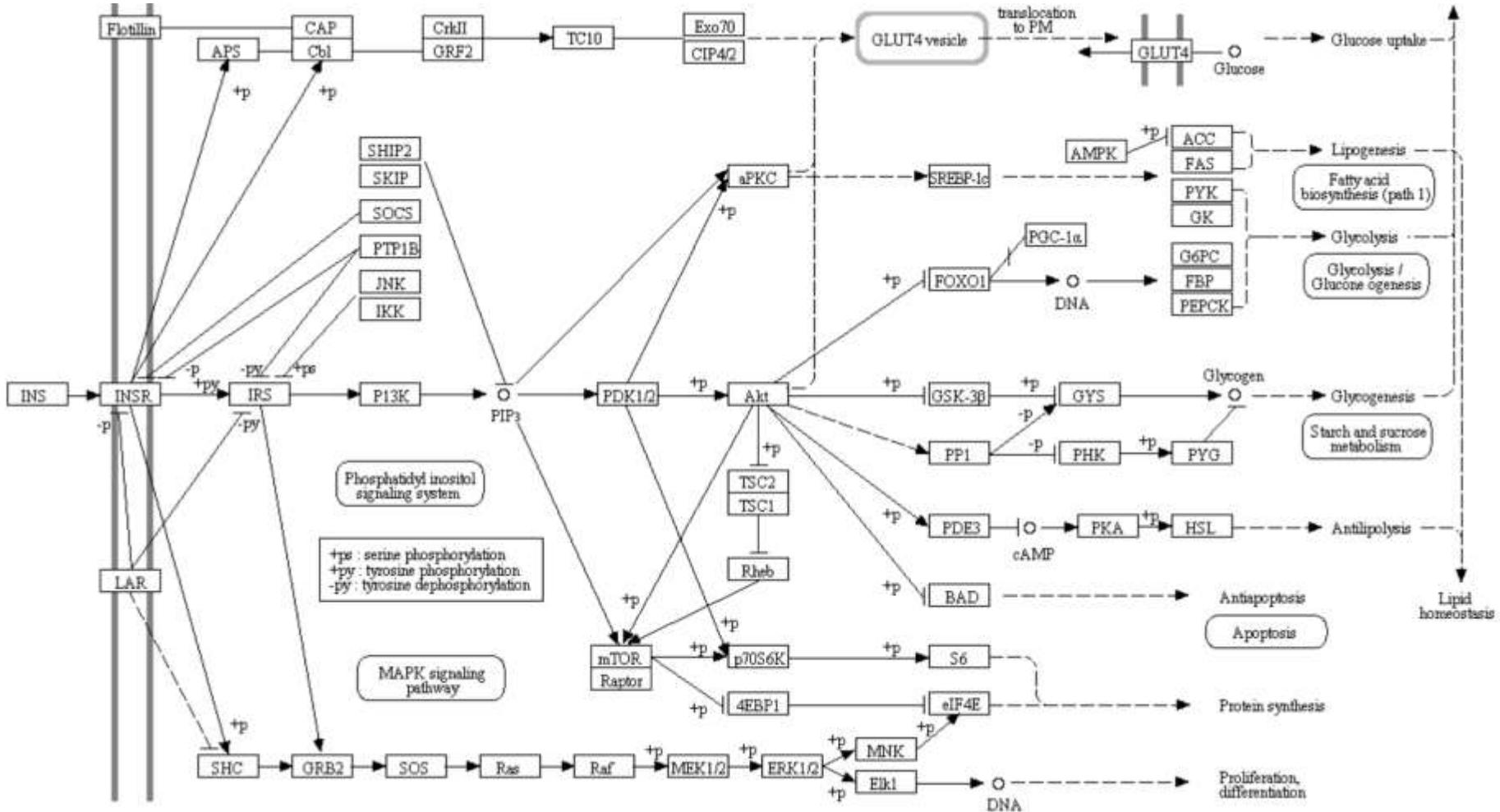
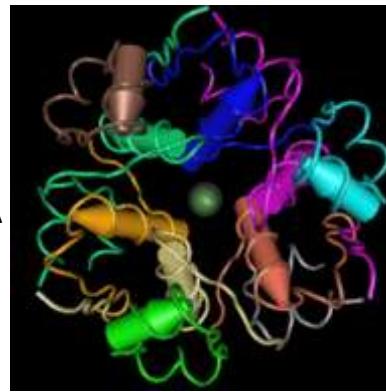
General protein information

NCBI Reference Sequences (RefSeq)

Related sequences

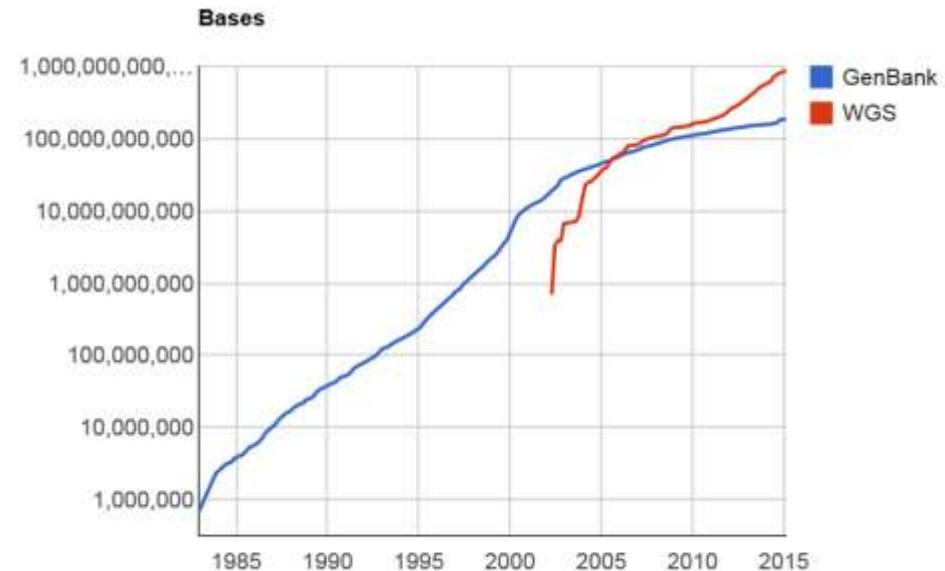
Additional links

>gi|631226408|ref|NP_001278826.1| insulin preproprotein [Homo sapiens]
 MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREA
 EDLQVGQVELGGGPGAGSLQPLALE GSLQKRGIVEQCCTSICSLYQLENYCN



The “Big” Data

- From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months



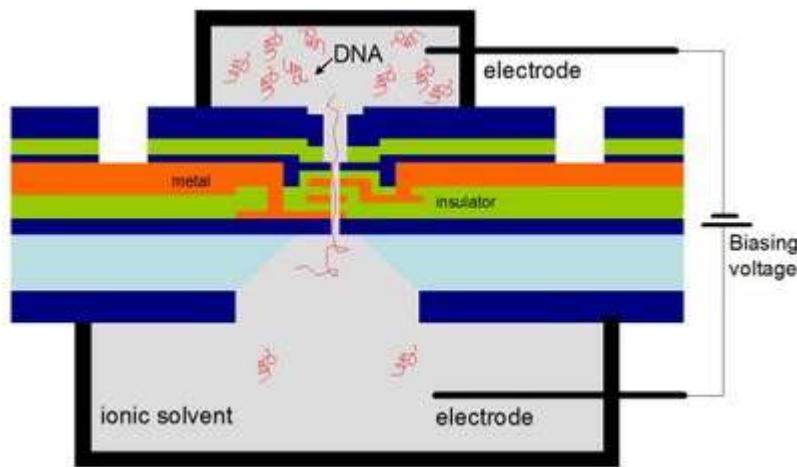
DNA Sequencer



©2010, Illumina Inc. All rights reserved.



IBM transistor sequencer



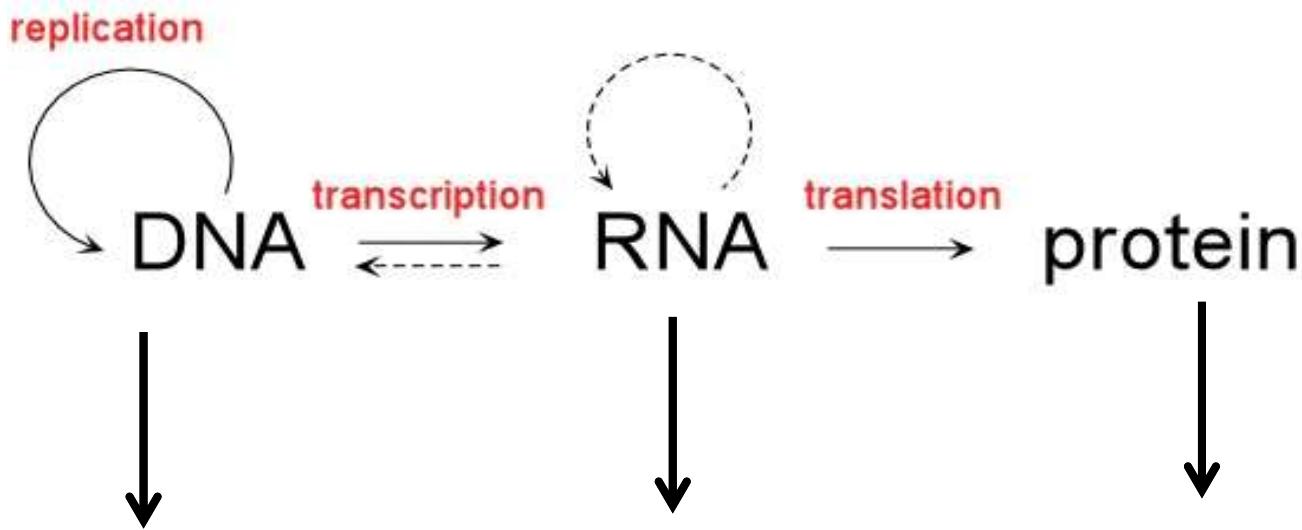
CryoEM



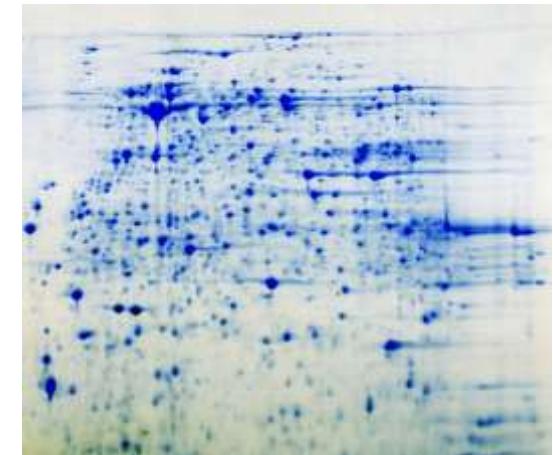
Mass Spec



The Central Dogma in 21st century

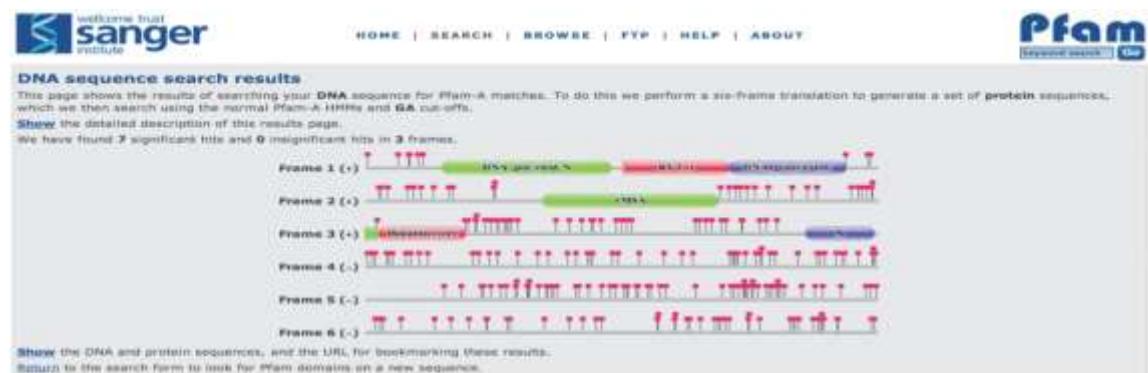
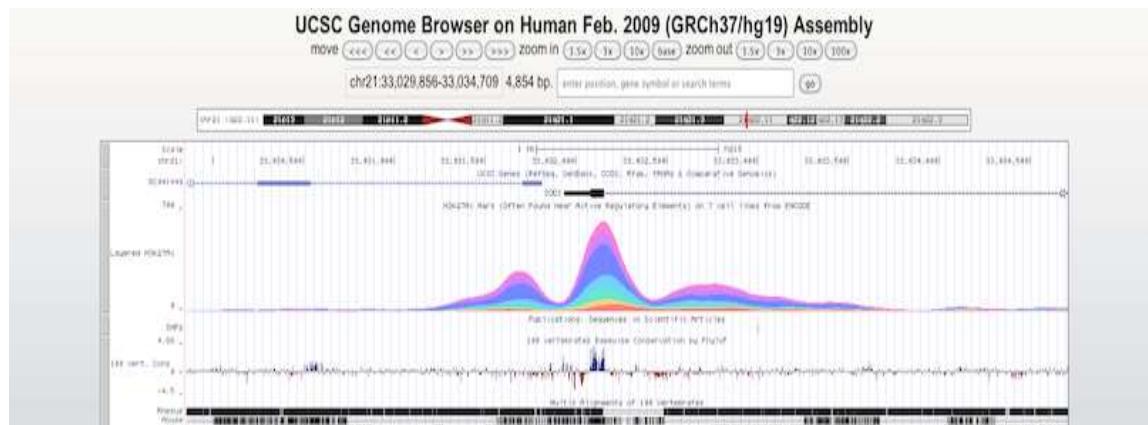


GGGGGTCGGAGTTCGAATTCTGGTAGAGGA
ACAGTCATTTGAAATCTGGCCTCGGGG
TACGTGTGCCTACTGAGTTCCCTGGAAACGGGA
GGTACGGGTGCCTACTGAGTTCCCTGGAAACGGGA
GGGTGAGAGCCCCGTCGGTAGAACACCCCC
GGTGCAGGGTTCCCTCGAGTTCCCTGGAAACGGGA
TTATGGTTCCCTGGAGTTCCCTGGAAACGGGA
ACCTTAATACATGGATAACGGTGGAAATGG
CCCTGGAAACGGGACGCCATAGAGGG
GGTGCGGGTTCCCTCGAGTTCCCTGGAAACGGGA
AGTTCCCTGGAAACGGGACGCCATAGAGGG
AATTGAAATCTGGCTCCCTGGGGCCCCGAG
AGGGTGAAGCCCCGTCGGTCGGAAACCCCC
CCCTGGAAATAGAGGGTGAAGAGCCCCGAGGG



Secondary database

- UCSC database
- GPCR database



Frame (series)	Family	Description	Entry type	Class	Envelope Start	End	Alignment Start	End	HMM From	To	HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
1 (+)	DNA_poly_nstat_N	DNA polymerase (viral) N-terminal domain	Family	info	189	514	183	514	1	379	379	373,7	1.7e-172	none	Show
1 (+)	RNCT_1	RNA-dependent RNA polymerase	Family	CL0022	538	762	538	762	2	214	214	306,6	2.7e-61	none	Show
1 (-)	DNA_poly_nstat_C	DNA polymerase (viral) C-terminal domain	Family	info	763	1009	763	1009	1	249	249	477,2	8.1e-144	none	Show
2 (-)	yMSA	Major surface antigen from hepatitis virus	Family	info	372	737	372	737	1	364	364	363,4	2.3e-109	none	Show
3 (-)	hep_core_N	Hepatitis core protein, putative zinc finger	Domain	info	1	24	1	24	4	27	27	56,7	2.5e-16	none	Show
3 (-)	hep_core_C	Hepatitis core antigen	Domain	info	28	209	28	209	1	187	187	315,3	1.1e-94	none	Show
3 (+)	X	Trans-activation protein X	Family	info	823	1064	923	1064	1	142	142	268,1	3.4e-80	none	Show

Comments or questions on the user interface may be sent to pfam@ebi.ac.uk. Our privacy policy.
The Wellcome Trust

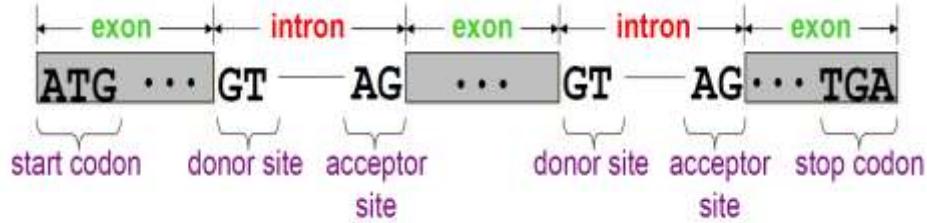
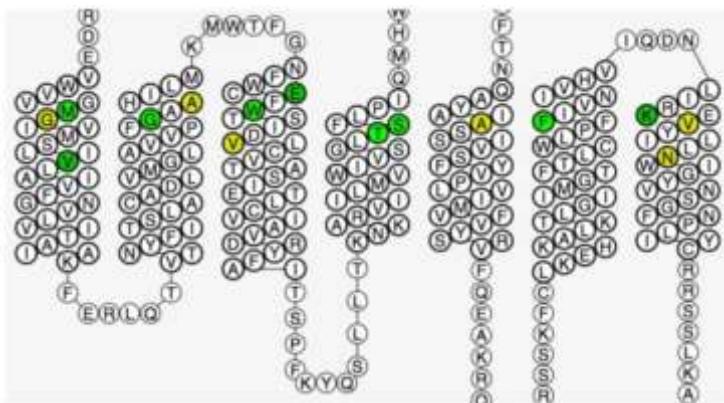
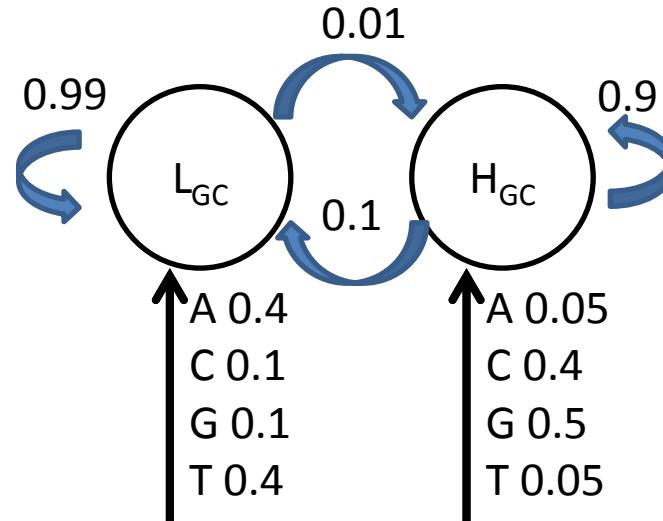
Database search “alignment”

- Longest Common Subsequences
- Smith-waterman algorithm
- heuristic search (BLAST, BLAT, Burrows-Wheeler Aligner, etc)

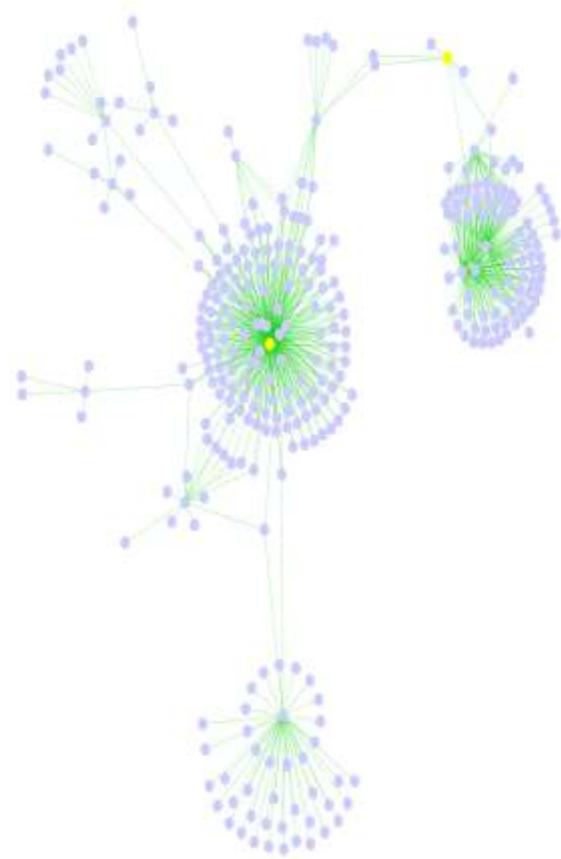
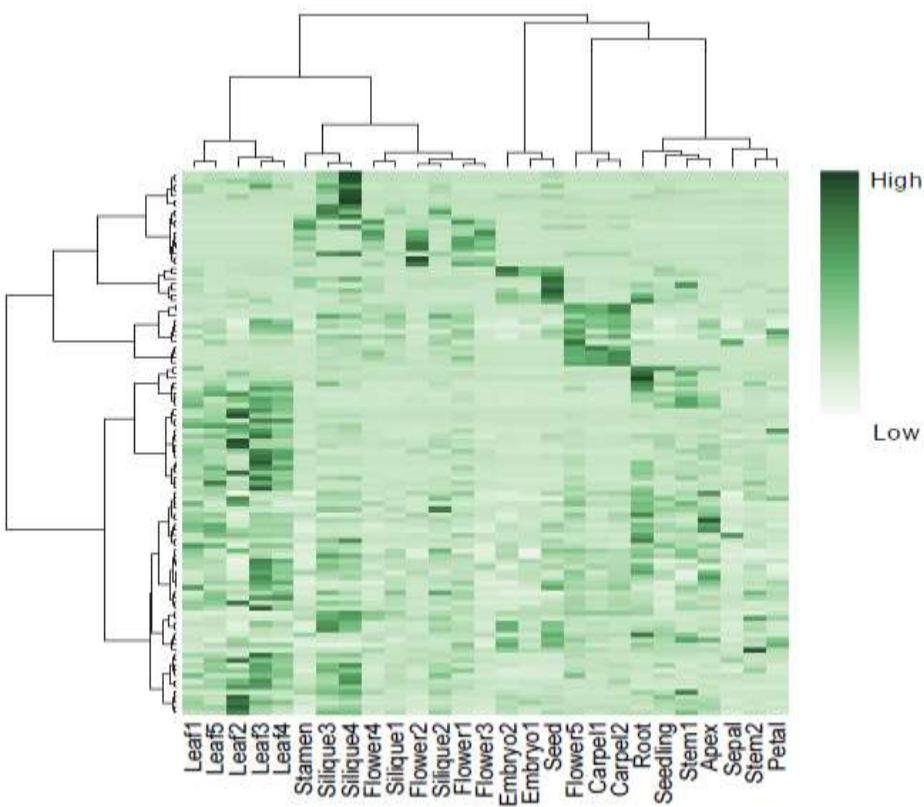
Sequence 1 = A--**CACACTA**
Sequence 2 = AG**CACAC**-A

Hidden Markov Model

- GenScan
 - Pfam/HMMER



Data mining in biological data



Human genetic study

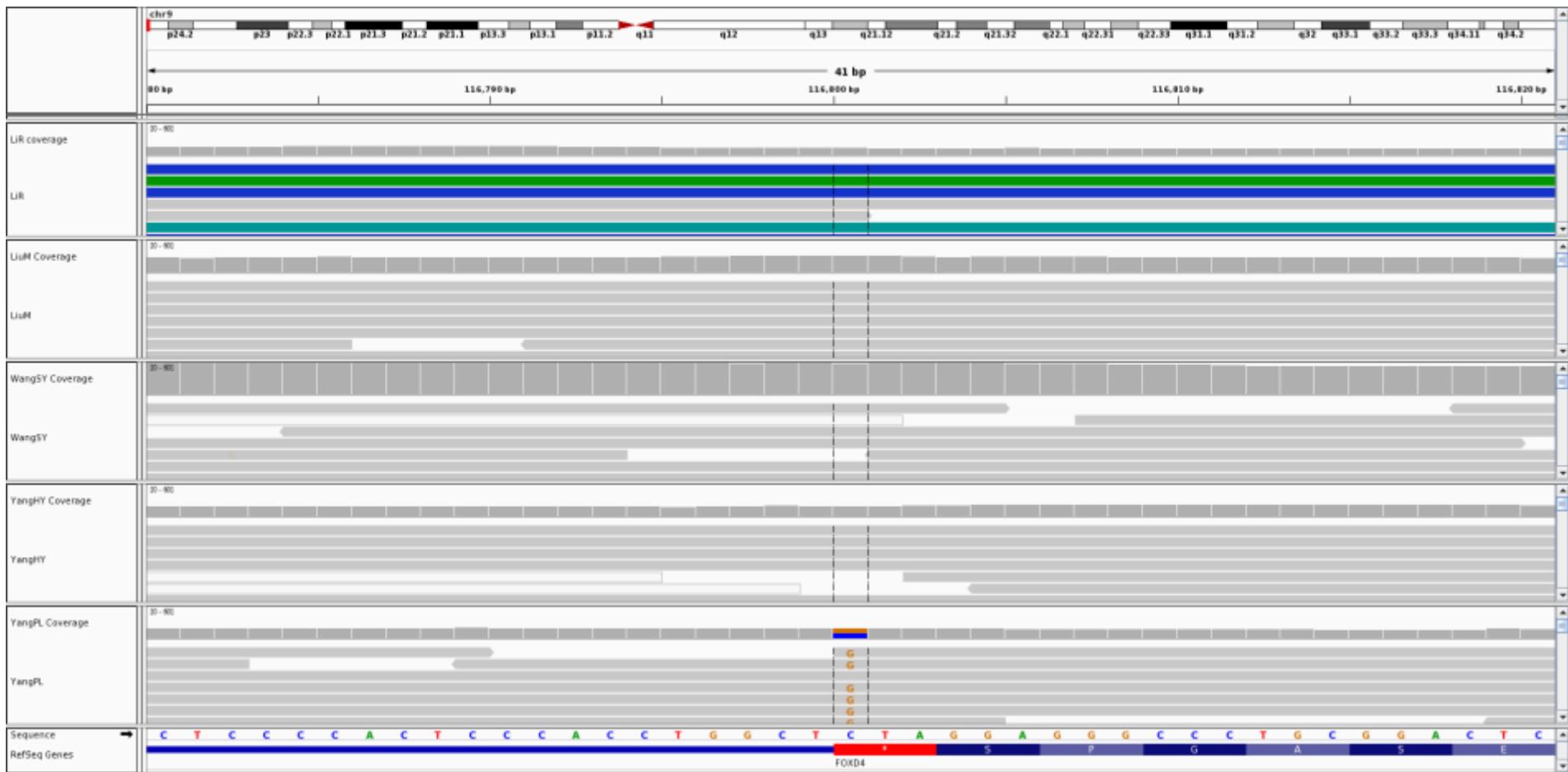
-- a case study

- Background
 - SNP (biomarker)
- Candidate-gene based study
 - HIV opportunistic infection
- Whole genome wide study
 - Hepatitis C

SNP (single nucleotide polymorphism)

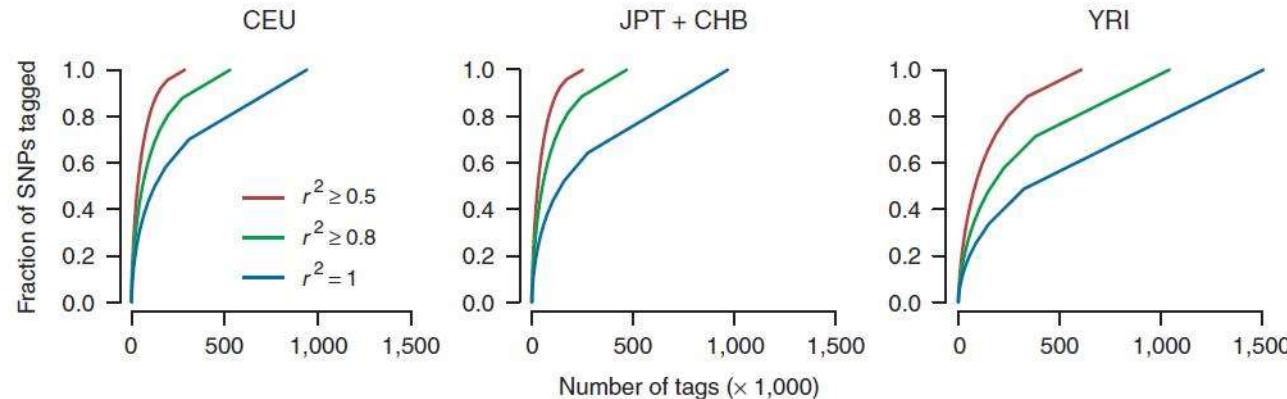
- Definition
 - DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered
- Common human variation
 - 11 million (MAF \geq 1%)
 - SNP frequency varies in different population
- 1000 genome project
 - Genotyped 25 population, ~2500 individuals
- Detected method
 - Sequencing
 - PCR-based methods
 - Chip (Illumina, Affy)
 - ...

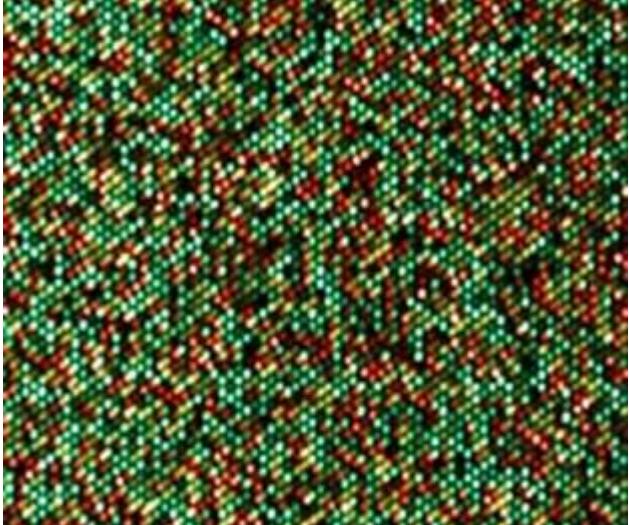
SNP detection



Genotyping platform on Chip

- Affy
 - Genome-wide Human SNP array 6
- Illumina
 - Human 1M-duo
- Coverage

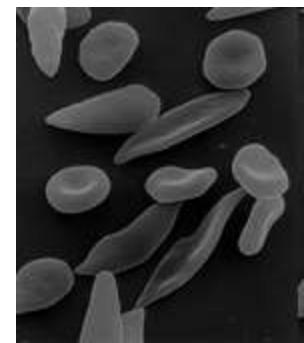




1

Association study in infection

- Infection disease exert evolution pressure in human population
 - Malaria and sickle-cell anaemia risk allele
- Advantage of association study
 - Linkage analysis need multiple affected and unaffected relatives
 - Family-based, case-control or cohort data
 - Fine localization and identification of causative loci with high-throughout technology



Association study methodology

- Chi-squared

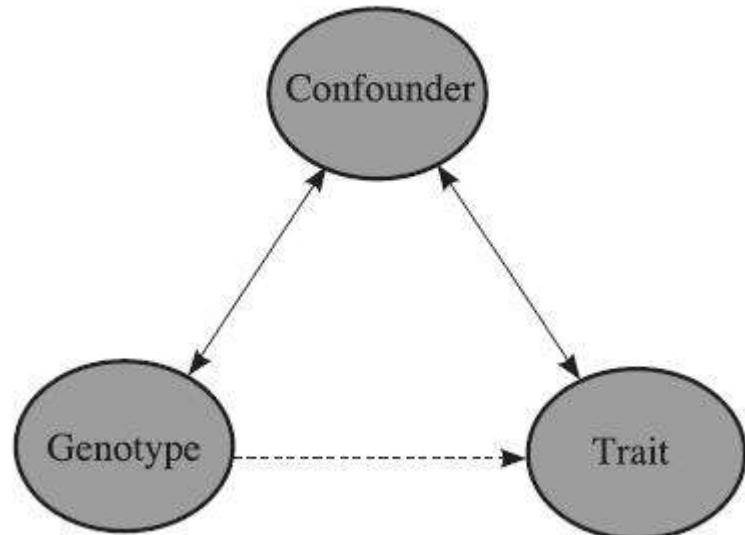
		Disease	
		Disease	Unaffected
Allele 1	Disease	p_{1D}	p_{1U}
	Unaffected	p_{2D}	p_{2U}

- Regression

$$g(E[\mathbf{y}]) = \mathbf{X}\boldsymbol{\beta}$$

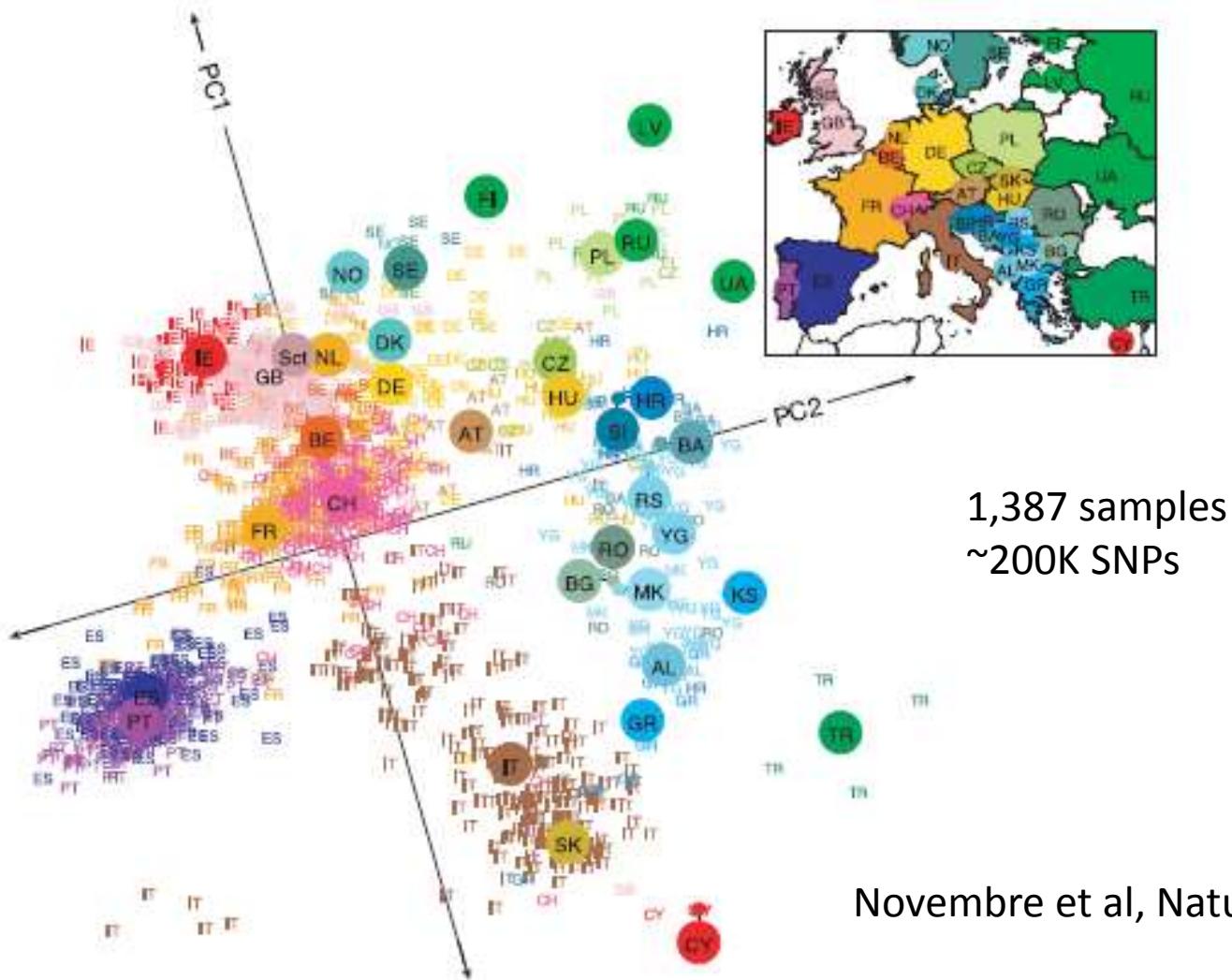
Association study methodology

- Confound factors



- Powers
 - $1-P(\text{false negative})$
 - Case-control study: genetic effect, Allele frequency ...

European population structure



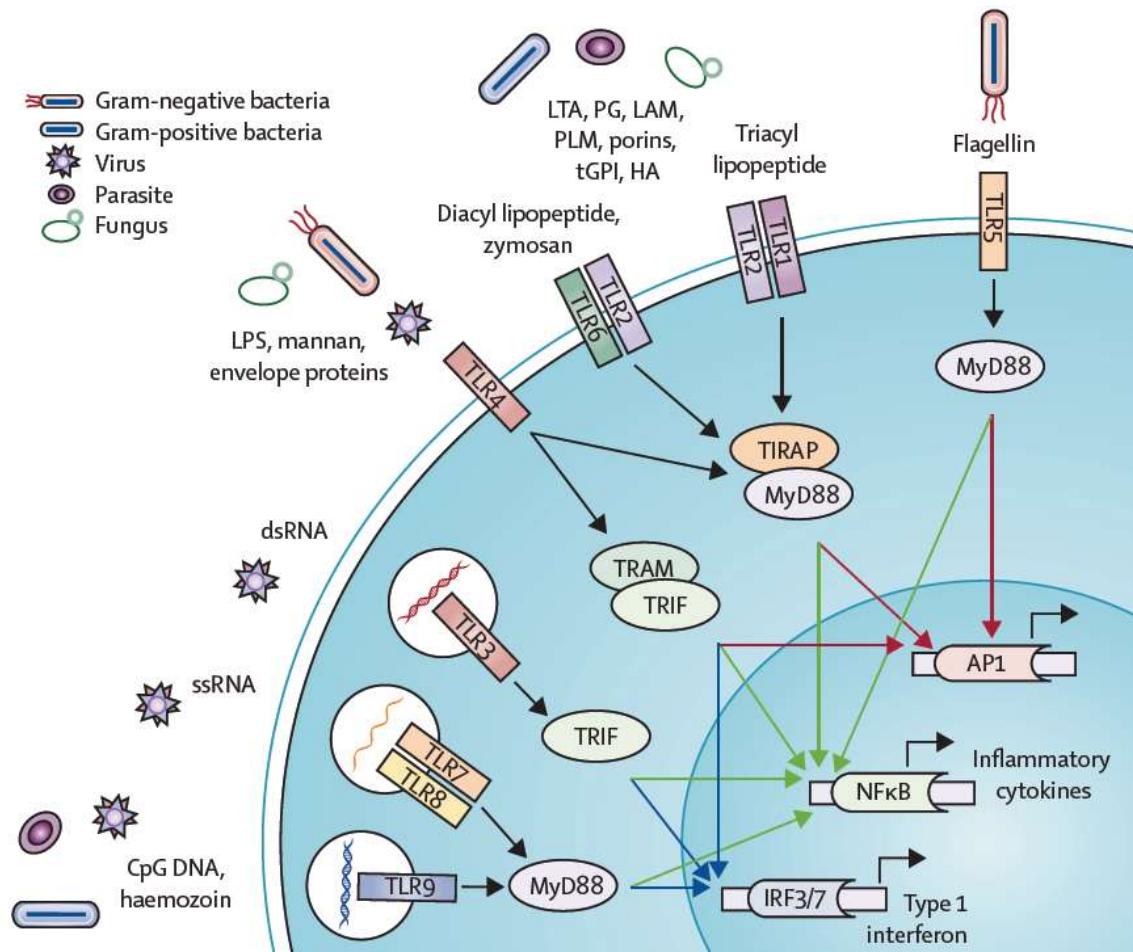
GWAS-basic analysis

- Quality control
 - MAF>0.05, HWE>0.001, GENO>0.95
 - Remove duplication or other mistakes
- Association analysis
 - Genetic model: Allelic (chisq 1df), Additive, Dominate, Recessive, Cochran-Armitage trend test, Genotypic test (chisq 2df)
 - QQplot and Manhattan plot
- Available software
 - Plink, GenABEL(R package) ...

GWAS-advanced analysis

- Population stratification
 - χ^2 divided by genomic inflation
 - IBS clustering in PLINK
 - PCA in EIGENSTRAT
- Imputation
 - MACH, IMPUTE...
 - MACH cutoff(>0.9) means free genotyping
- Meta-analysis
 - Reverse variance pooling method
 - Carefully prepare the data (same population, same reference allele, same phenotype unit, etc)

TLR4 SNPs association study in HIV opportunistic infection



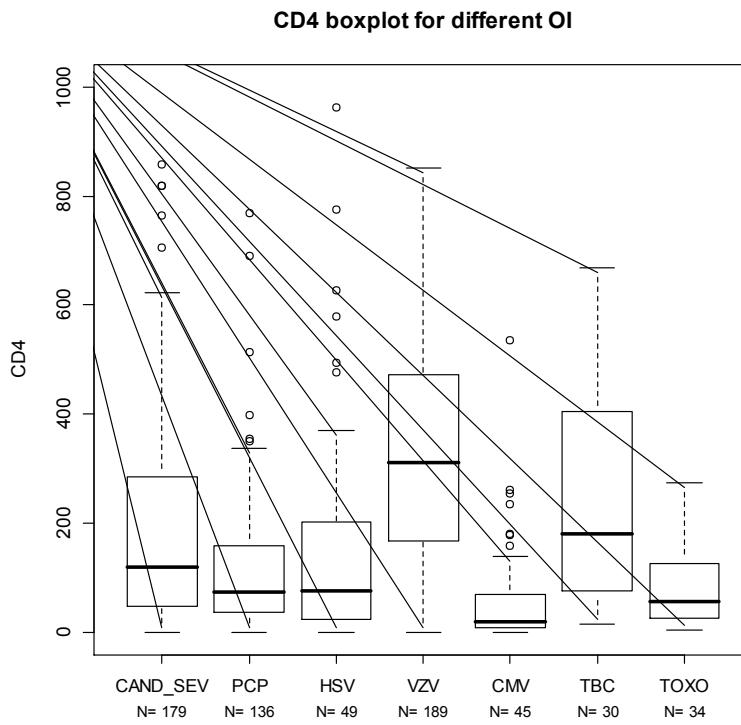
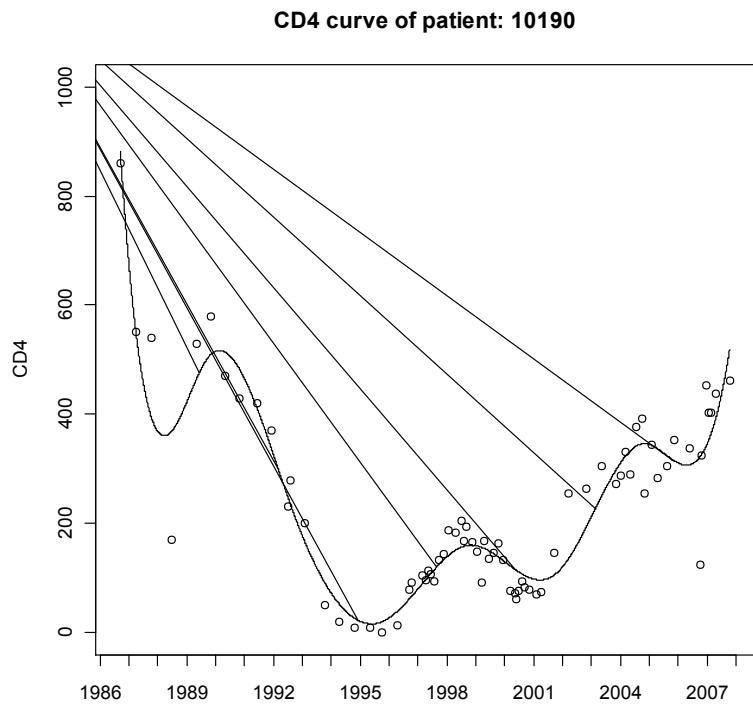
TLR (Toll-like receptors) history

- Toll: function in the embryonic dorsal-ventral development of *Drosophila* (1988, *cell* 52:269)
- *Drosophila* with a loss of function mutation for Toll exhibits a high susceptibility to fungal infection (1996, *cell* 86:973)
- TLR: the so-called Toll-like receptors, human homolog genes for Toll(1997, 1998)
- TLR4 is the LPS sensor in both mice and humans (1998, *Science* 282:2085)
- Inflammatory caspases are innate immune receptors for intracellular LPS (2014, *Nature* 514:187)

The TLR4 D299G SNP Influences Susceptibility to Opportunistic Infections in the Swiss HIV Cohort Study

- 1585 Caucasian patients are included from SHCS
- Poisson regression used to detect association
 - Neutral model
 - Additive model
 - adjusted by cofactor such as age, sex, infection risk factors and year of SHCS entry
- OIs
 - Fungal infection
 - severe candidiasis (mainly candida oesophagitis)
 - *Pneumocystis jirovecii* pneumonia (PCP)
 - Viral infection
 - HSV infection (mucocutaneous ulceration or HSV disease)
 - VZV infection (e.g. multidermatoma or relapsing zona)
 - CMV infection (CMV disease or retinitis)
 - Mycobacterium infection
 - tuberculosis
 - Parasite infection
 - toxoplasmosis
- The permutation false discovery rate (FDR)
 - the genotyped SNPs in all the patients are randomly shuffled, and then the same poisson regression is done. We take the ratio of the cases in 1000 times shuffle in which random pvalue is less than the real one as Qvalue

CD4 distribution of OIs



Neutral model

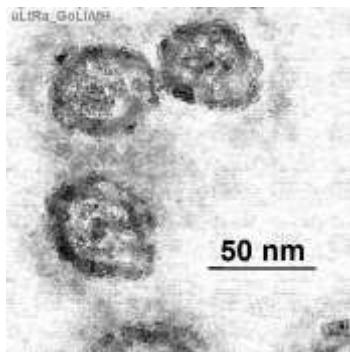
Ols	TLR_D299G	CD4+ below		Days at risk	Case of OI	Others	IR (per year)		IRR	Pvalue
								(per year)		
CAND_SEV	0/0	300	1021865	97	711	0.0347	-	-	-	-
	0/1	300	96420	11	72	0.0417	1.2018 (1.652 0.874)	0.563		
	1/1	300	3953	0	2	0	0 (Inf 0)	1		
PCP	0/0	200	477977	54	403	0.0413	-	-	-	-
	0/1	200	49465	11	37	0.0812	1.9684 (2.74 1.414)	0.041		
	1/1	200	2662	0	2	0	0 (Inf 0)	1		
HSV	0/0	200	477977	29	428	0.0222	-	-	-	-
	0/1	200	49465	4	44	0.0295	1.3328 (2.272 0.782)	0.59		
	1/1	200	2662	0	2	0	0 (Inf 0)	1		
VZV	0/0	400	1841694	82	1083	0.0163	-	-	-	-
	0/1	400	190490	12	113	0.023	1.4149 (1.927 1.039)	0.262		
	1/1	400	4419	0	2	0	0 (Inf 0)	1		
CMV	0/0	100	193765	32	239	0.0603	-	-	-	-
	0/1	100	21441	5	20	0.0852	1.4121 (2.284 0.873)	0.473		
	1/1	100	0	0	0	NaN	-	-	-	-
TBC	0/0	400	1841694	12	1153	0.0024	-	-	-	-
	0/1	400	190490	4	121	0.0077	3.2227 (5.741 1.809)	0.043		
	1/1	400	4419	0	2	0	0 (Inf 0)	1		
TOXO	0/0	200	477977	21	436	0.016	-	-	-	-
	0/1	200	49465	6	42	0.0443	2.7608 (4.386 1.738)	0.028		
	1/1	200	2662	0	2	0	0 (Inf 0)	1		

Incidence of OIs under immune suppression by TLR4 SNP

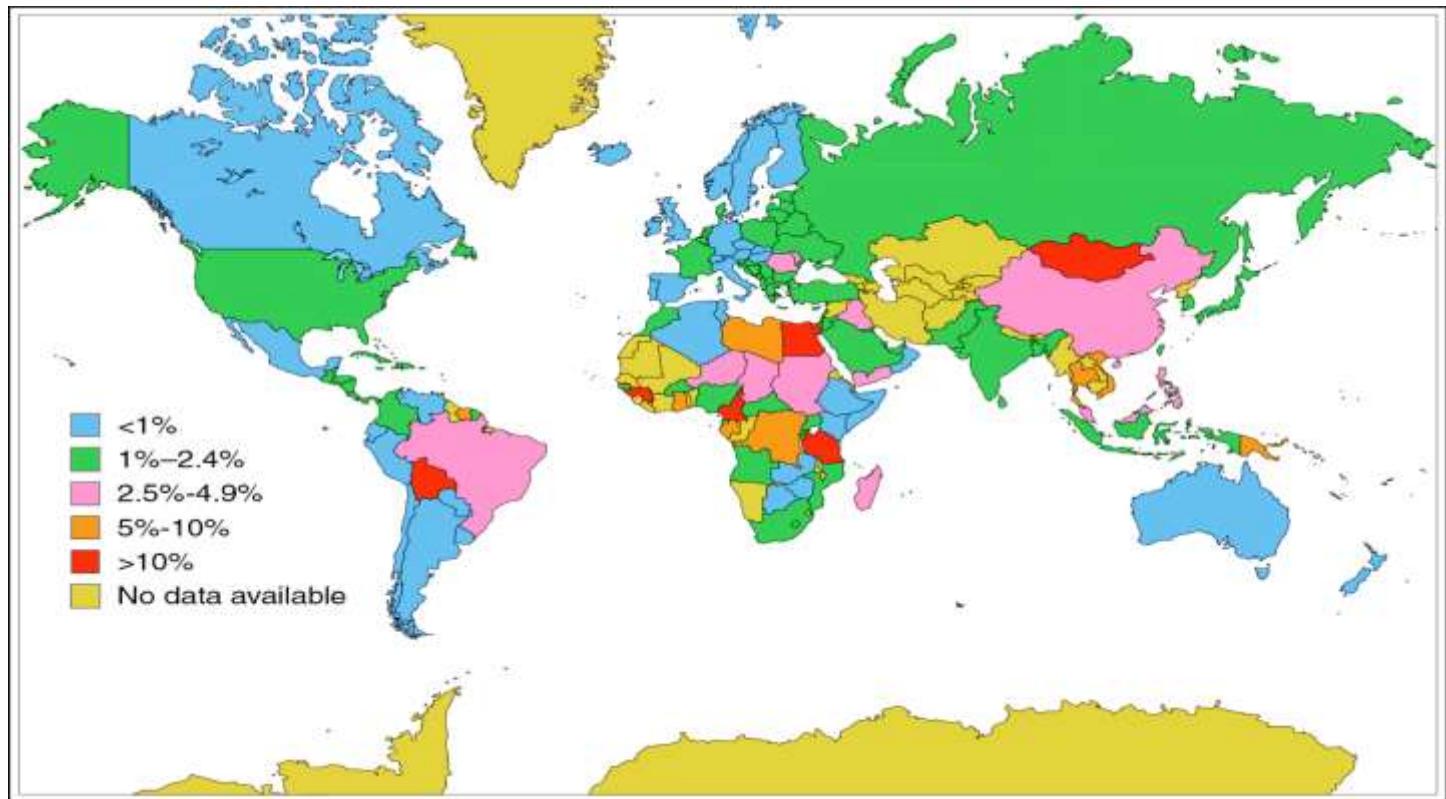
Additive model

OIs	CD4 cutoff	Incidence Rate Ratio	95% CI	Pvalue	Qvalue
CAND_SEV	300	1.1	0.8-1.5	0.7	0.7
PCP	200	2.0	1.4-2.7	0.047	0.040
HSV	200	1.3	0.8-2.2	0.6	0.6
VZV	400	1.4	1.0-1.9	0.3	0.3
CMV	100	1.4	0.9-2.2	0.3	0.3
TBC	400	2.6	1.5-4.3	0.077	0.057
TOXO	200	2.4	1.6-3.7	0.041	0.031

Genome wide association study in Hepatitis C



- █ < 1%
- █ 1.0-2.4%
- █ 2.5-4.9%
- █ 5-10%
- █ >10%
- █ NA



Method -- clinic

- Chronic HCV infection
 - anti-HCV seropositivity (using ELISA/RIBA) and detectable HCV RNA by quantitative assays
- Spontaneous clearance
 - HCV-seropositivity and undetectable HCV RNA in patients without previous antiviral treatment
- Response to treatment
 - at least 80% of the recommended dose PEG-IFN /RBV during the first 12 weeks
 - Sustained viral response (SVR)
 - undetectable HCV RNA in serum >24 weeks after treatment termination
 - Non-response (NR)
 - Others

Method -- genotyping

- Illumina 1M-Duo chip for SCCS
- Illumina Humanhap650-Quad beadchips for SHCS study (including part of work using Illumina Humanhap550)
- Illumina Beadstudio software used for genotype calling

Method --association analysis

- Quality control
 - MAF>0.01, HWE>0.001, GENO>0.95, mind>0.95
 - Remove duplication and other cryptic relatedness
- Basic association analysis
 - Allelic based analysis or Cochran-Armitage trend test
 - Logistics regression considering covariates
 - Significance cutoff 5E-8
 - QQplot and Manhattan plot
- Applied software
 - Plink and Haplovie

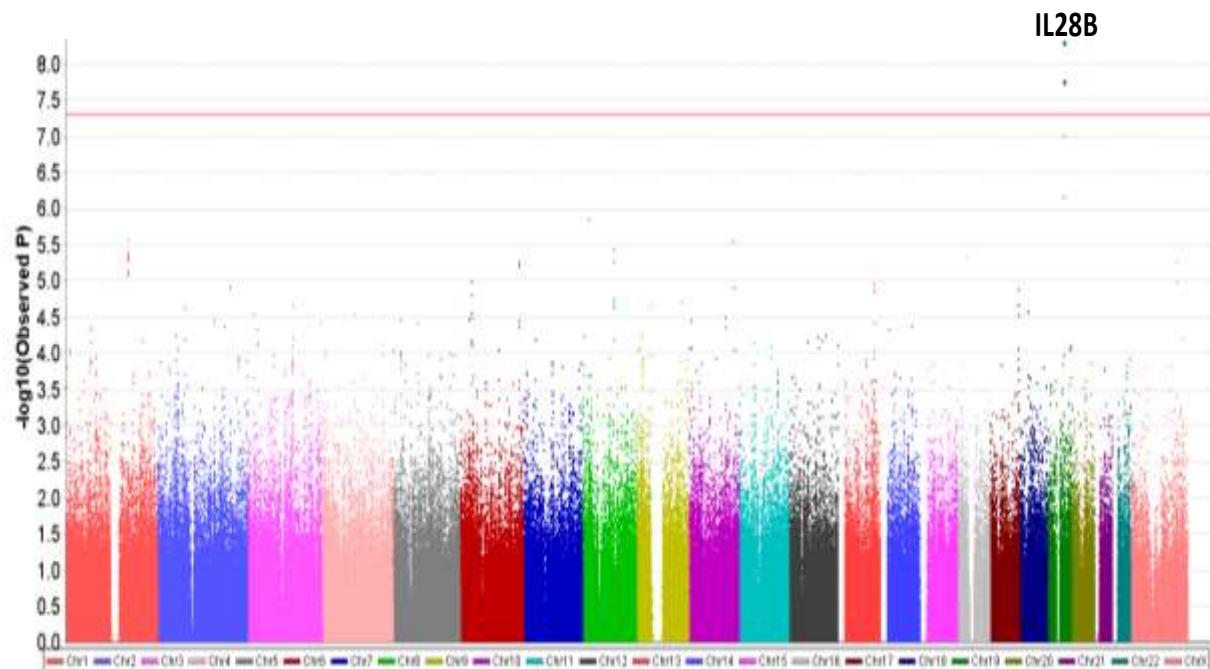
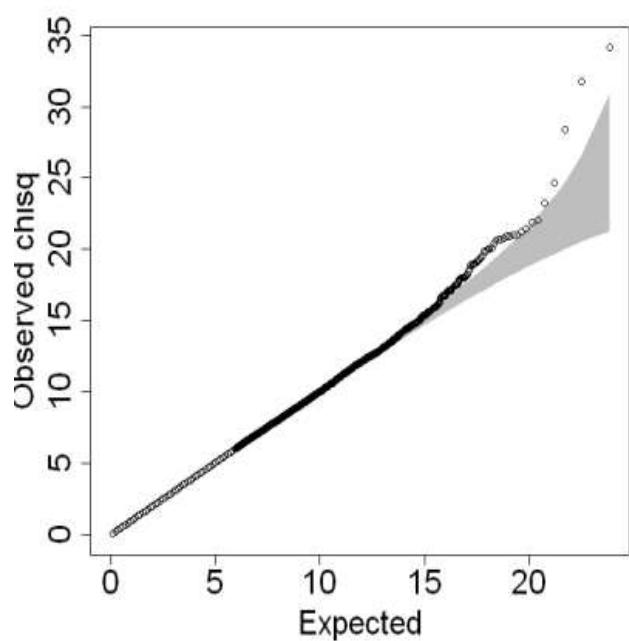
Spontaneous Clearance demographic table

Characteristics (N, proportion)	SHCS			SCCS		
	Chronic Infection	Spontaneous Clearance	P	Chronic Infection	Spontaneous Clearance	P
N	201	199		828	87 (+73 DE)	
Age (median, IQR)	33.75 (8.93)	33.86 (9.65)	0.7	44.15 (14.03)	37.47 (8.59)	<0.001
Male sex	105 (52.2%)	136 (68.3%)	0.001	516 (62.3%)	48 (55.2%)	0.2
HBV antigen positive	21 (10.4%)	8 (4%)	0.01	8 (1%)	4 (4.6%)	0.03
Log HCV RNA (median, IQR)	6.086 (1.346)			5.877 (0.993)		
HCV genotypes						
1	78 (39.2%)			396 (47.8%)		
2	5 (2.5%)			83 (10%)		
3	60 (30.2%)			240 (29%)		
4	22 (11.1%)			70 (8.5%)		
Other/unknown	36 (18%)			39 (4.7%)		

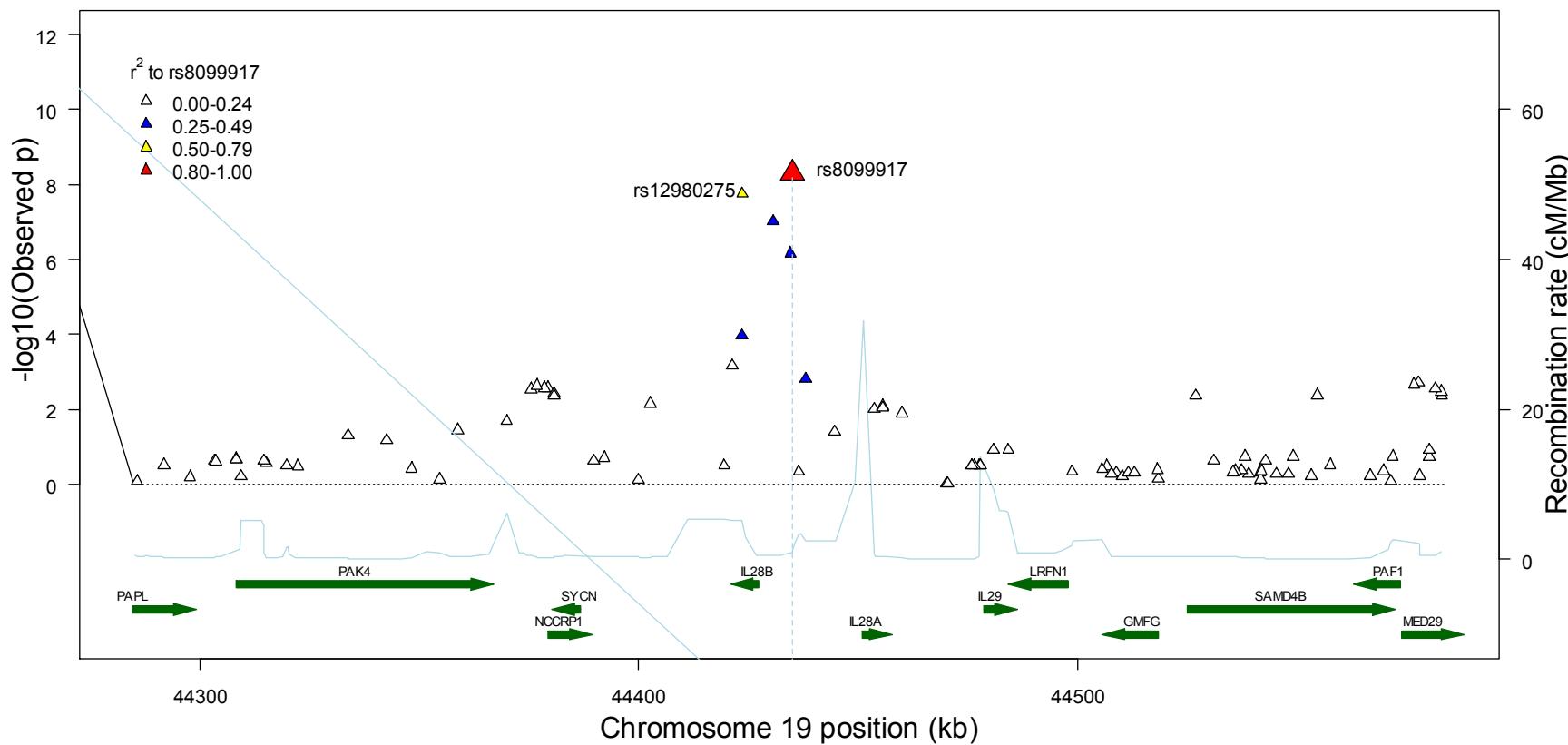
Response to Treatment demographic table

Characteristics (N, proportion)	NR	SVR	P
N	174	315	
Age (median, IQR)	19 (9)	20 (9)	0.04
Male sex	119 (68.4%)	185 (58.7%)	0.04
HBV antigen positive	2 (1.1%)	3 (1%)	0.9
Log HCV RNA (median, IQR)	5.964 (0.844)	5.835 (1.226)	<0.001
HCV genotypes			
1	105 (60.3%)	94 (29.8%)	Ref
2	8 (4.6%)	53 (16.8%)	<0.001
3	29 (16.7%)	142 (45.1%)	<0.001
4	19 (10.9%)	17 (5.4%)	1
Other/unknown	13 (7.5%)	9 (2.9%)	0.6
Heavy drinker	31 (17.8%)	35 (11.1%)	0.03
Liver biopsy			
Inflammation	23 (13.2%)	45 (14.3%)	0.5
steatosis	85 (48.9%)	150 (47.6%)	0.5
Severe fibrosis	55 (31.6%)	59 (18.7%)	0.003

IL28B identification in the GWA for response to treatment



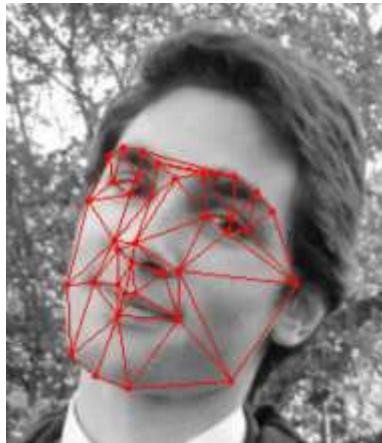
Region association plot of IL28B





未來 · 已來

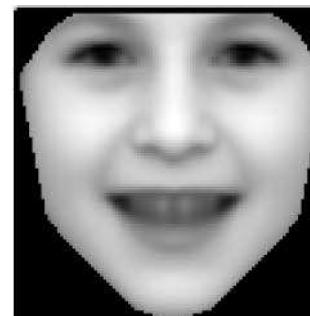
Diagnostically relevant facial gestalt information from ordinary photos



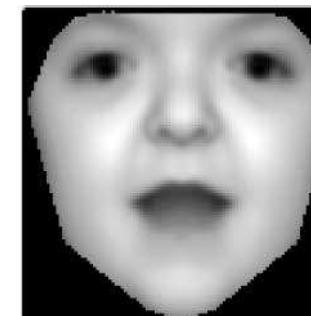
Elife, 2014



control



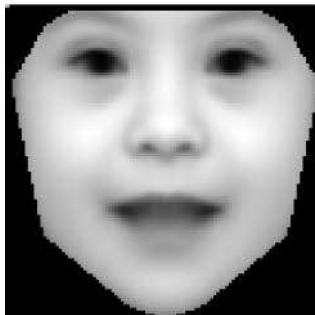
Angelman



Apert



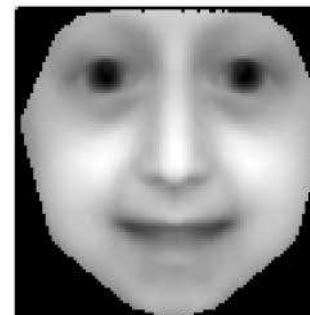
Cornelia de Lange



Down



Fragile X



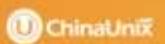
Progeria



Treacher-Collins



Williams-Beuren



THANKS