

MASSA

MultiAgent System for SNP Annotation

Topics

- * MASSA Overview
 - * Functions and Databases
 - * MultiAgent Schema
- * Performance
 - * Benchmarking
- * Applications of Annotation and Enrichment Analysis
 - * Principal Component “AIMs”
 - * Admixture Clusters AIMs
 - * Illumina Arrays

MASSA OVERVIEW

MASSA Overview

Functions



EpiGen-Brasil

- * Annotation
 - * SNP-based
 - * dbSNP, UCSC, GWAS NHGRI
 - * Gene-based
 - * GO, PharmGKB, OMIM, HGNC
- * Summary
 - * Annotation terms count (eg. Transcript regions, Disorders)
- * Gene Enrichment Analysis
 - * Fisher's Exact Test
 - * Hypergeometric Test (future implementation)



Current Databases

Database	Info
dbSNP	Genetic variants info
UCSC	Genomic info (SNPs annotation and conservation scores)
GO	Gene annotation (molecular function, cell component, biological process)
OMIM	Diseases database
HUGO	Gene nomenclature
PharmGKB	Pharmacogenetics info (Pathways, diseases, drugs)
GWAS NHGRI	Catalog of GWAS Studies

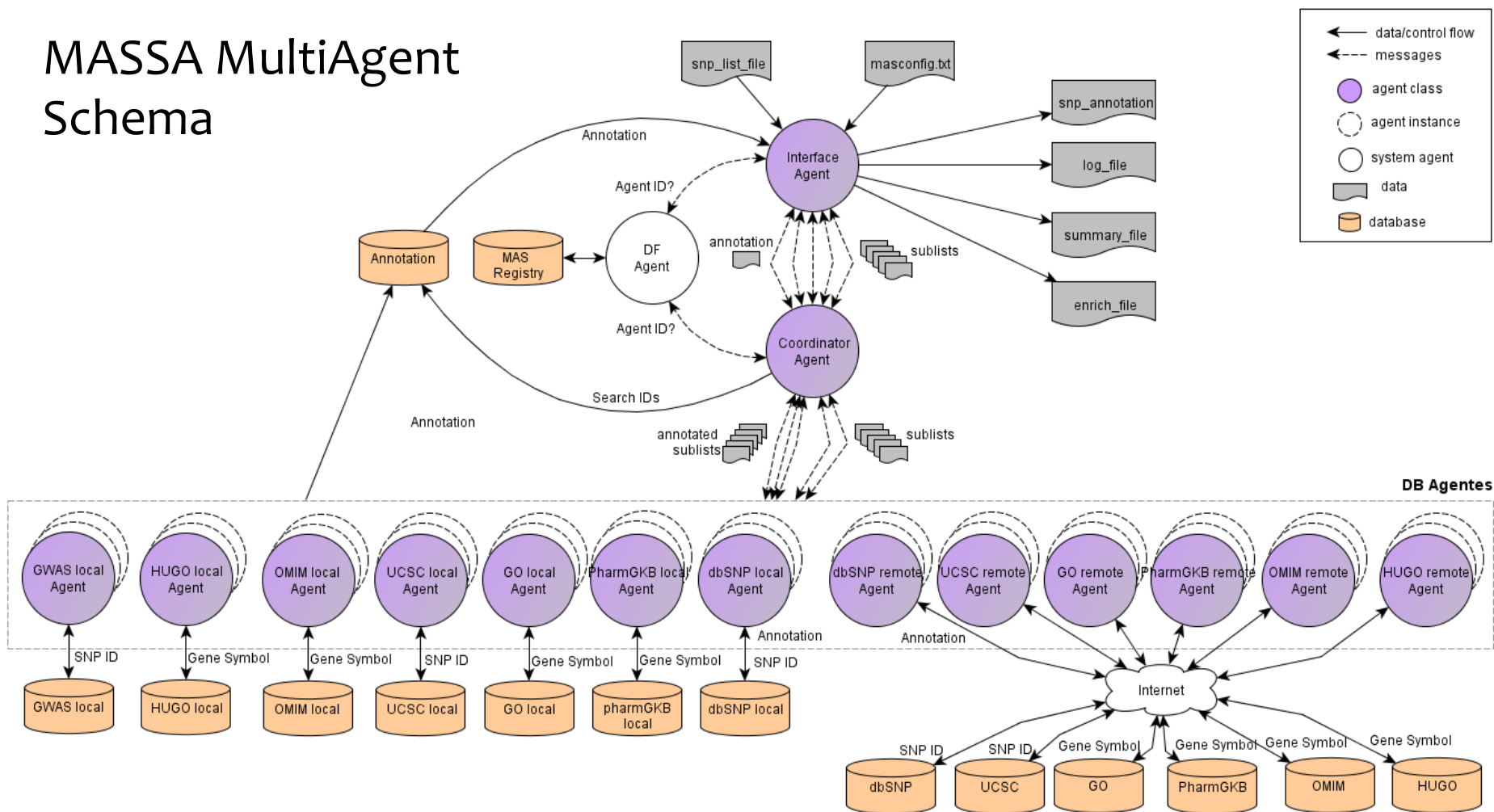
MASSA Overview Databases



Current Attributes

1.Polymorphism Id	11.Assembly Build Version	21.Polymorphism Class	31.CytoLocation	41.Gene Name	51.p-value
2.Polymorphism Type	12.Assembly Coord Start	22.Functional Class	32.Gene Status	42.Gene Synonyms	52. Disease/Trait
3.Gene Symbol	13.Assembly Coord End	23.Conservation Scores	33.Gene Map Methods	43.Locus Type	53.Sample/ Populations
4.Gene ID	14.mRNA accession	24.Pathway	34.Disorders	44.Locus Group	dbSNP
5.Transcript Region	15.mRNA version	25.Drugs	35.MIM Ids	45.Gene Family Tag	UCSC
6.Gene Coord	16.Alleles	26.Diseases	36.Inheritance	46. Gene Family	GWAS NHGRI
7.Chromosome	17.Frequency	27.Related Genes	37.PhenMap Methods	47.Pubmed	PharmGKB
8.Chromosome Position	18.Strand	28.Molecular Function	38.Comments	48.Reported Genes	Gene Ontology
9.Ancestral Allele	19.RefUCSC	29.Cellular Component	39.HGNC ID	49.Strongest SNP Risk Allele	OMIM
10.Orientation	20.ObservedUCSC	30.Biological Process	40.Gene Symbol	50.Context	HGNC

MASSA MultiAgent Schema



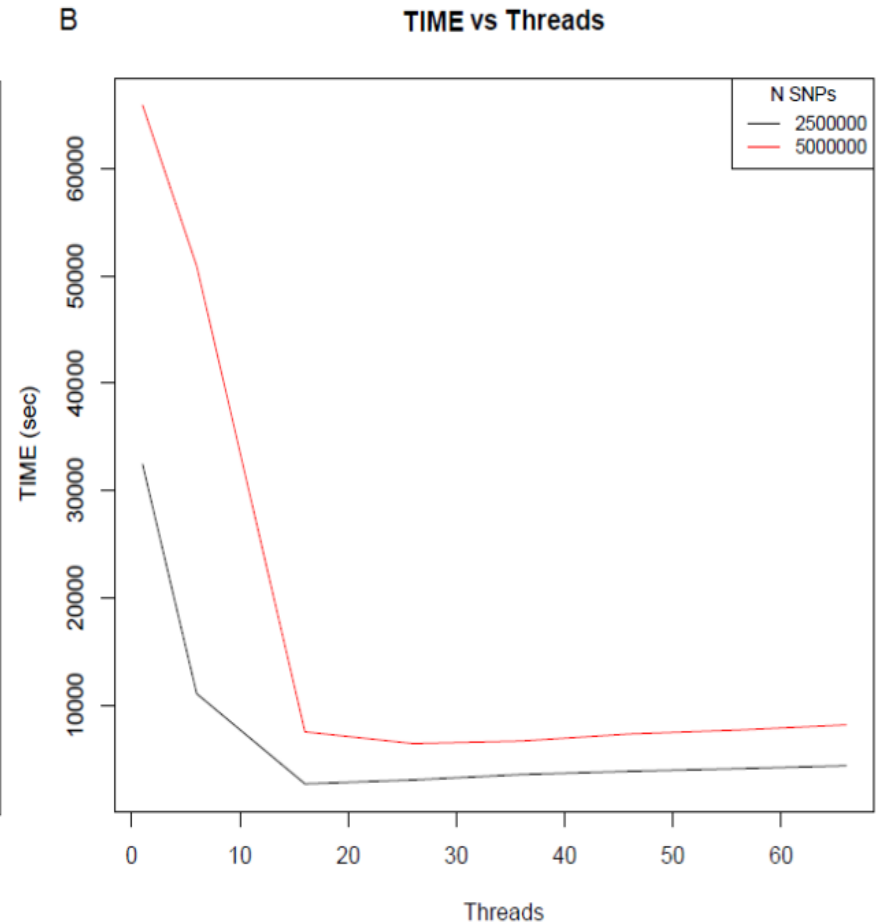
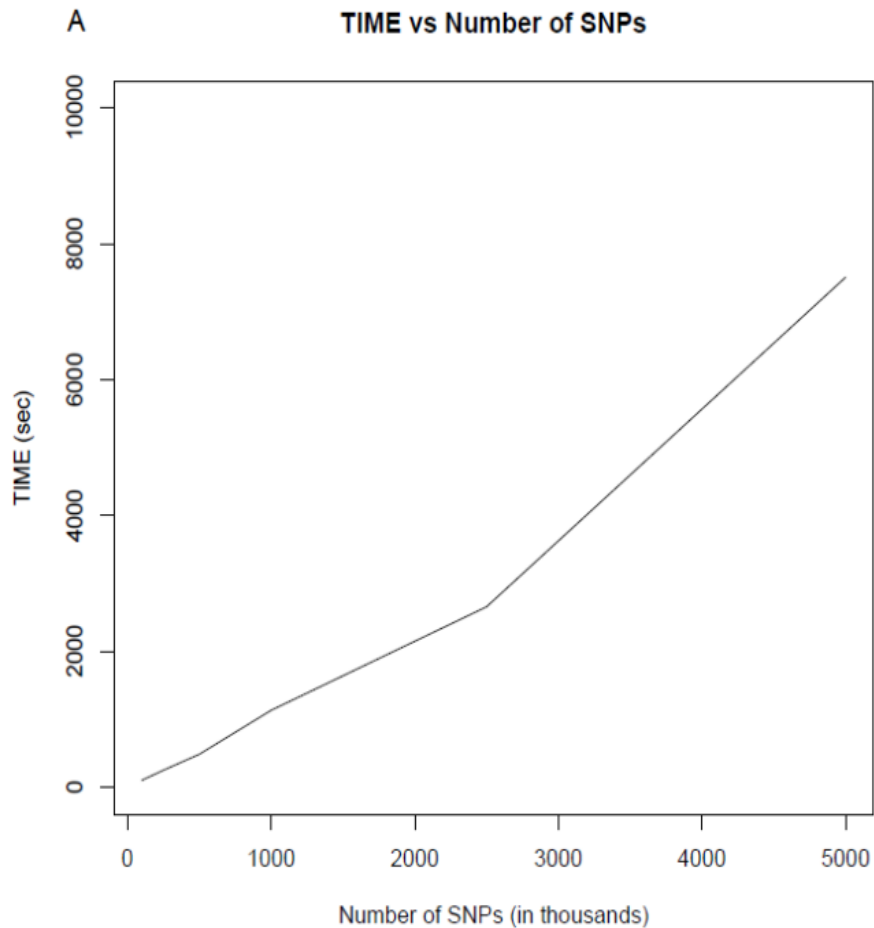
PERFORMANCE

Performance Benchmarking



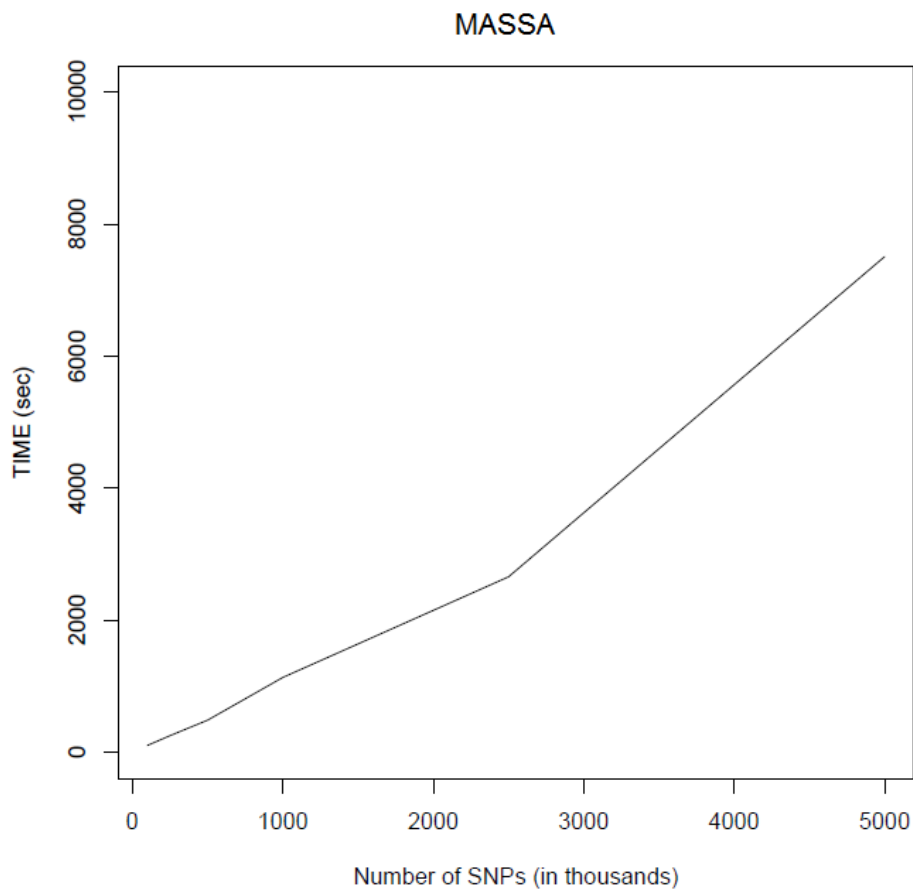
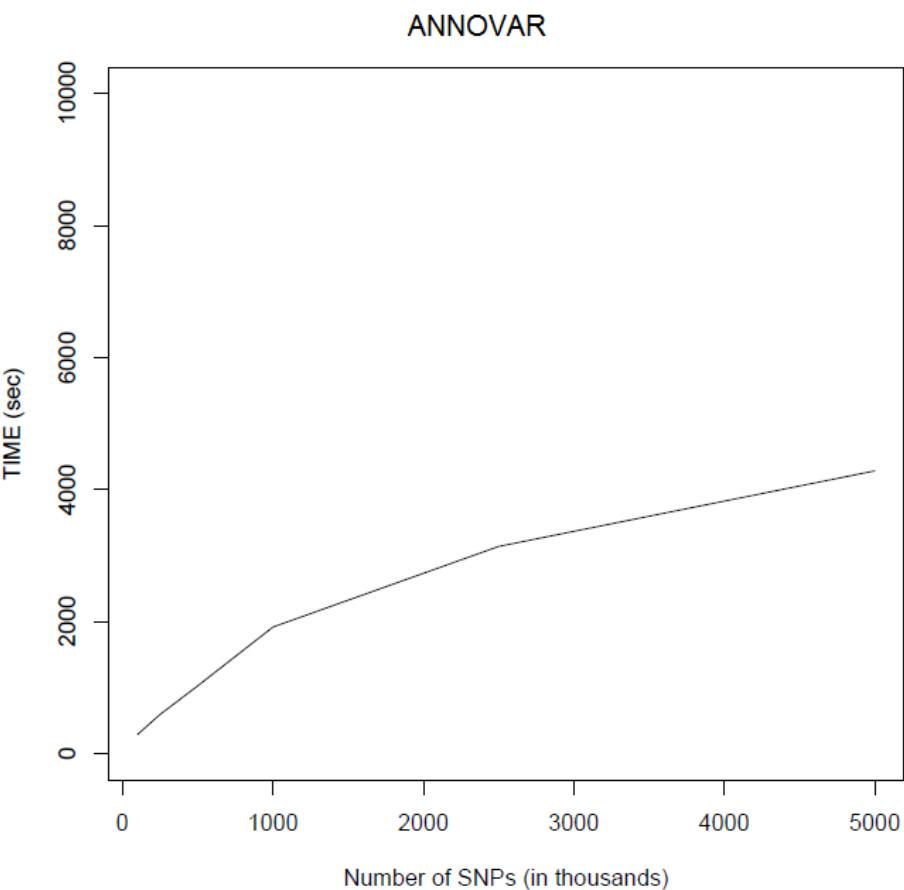
EpiGen-Brasil

MASSA Benchmarking



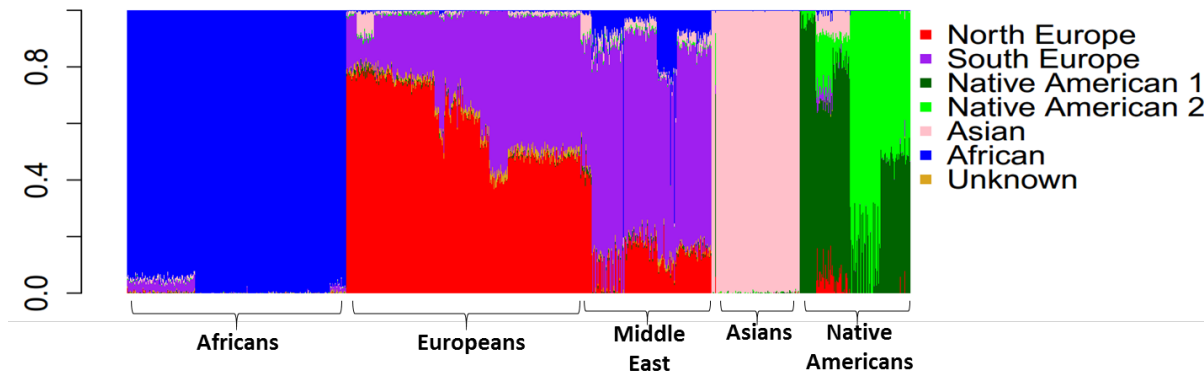


Comparison of MASSA and ANNOVAR



APPLICATIONS

* Admixture Analysis K=7



- * Selected SNPs with allele frequency differences higher than 0.3 between a given cluster and the others

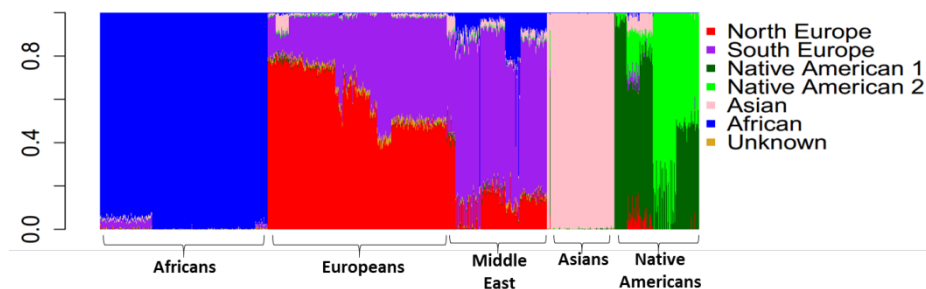
$$\boxed{?} \text{abs}(\text{Cluster1 } fa - \text{mean}(\text{Cluster2...7 } fa)) = \delta > 0.3$$

Applications

PCA “AIMs”



- * SNPs selected per K
 - * K1 – 7409 (1872 genes)
 - * K2 – 27821 (4449 genes)
 - * K3 – 15748 (3159 genes)
 - * K4 – 10655 (2369 genes)
 - * K5 – 7473 (1993 genes)
 - * K6 – 6321 (1757 genes)
 - * K7 – 6538 (1688 genes)





Applications
PCA “AIMs”
Annotation

[illegible]



EpiGen-Brasil

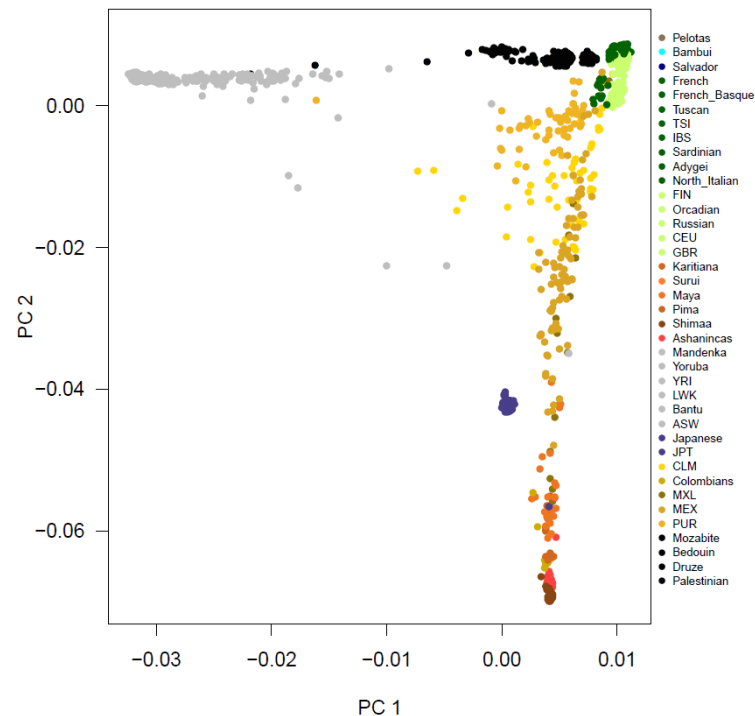
Applications Illumina Arrays Summary

Transcript Region (dbSNP)	2.5M	5M
Intergenic	1.194.986	2.137.587
UTR-5	3760	9055
UTR-3	25311	59543
nearGene-5	7011	12161
nearGene-3	2174	3743
splice-5	50	180
splice-3	62	220
Intron	842.570	1.606.930
cds-reference	51604	115498
cds-synon	15480	30803
cds-indel	0	2
Missense	13533	31519
Frameshift	2	7
STOP-GAIN	184	566
STOP-LOSS	7	19
ncRNA	7564	19062

Applications Illumina Arrays Summary

Gene	2.5M (SNPs)	Gene	5M (SNPs)
CSMD1	4811	CSMD1	8008
RBFOX1	3224	RBFOX1	5495
PTPRD	2963	PTPRD	5211
WVOX	2395	WVOX	3990
CDH13	2243	CDH13	3715
FHIT	2031	CNTNAP2	3579
CNTNAP2	1961	FHIT	3462
DMD	1865	DMD	3233
LRP1B	1777	MACROD2	3124
MACROD2	1687	LRP1B	3121
PARK2	1503	DLG2	2805
DLG2	1458	ROBO2	2634
ROBO2	1410	SGCZ	2589
SGCZ	1404	PARK2	2568
MAGI2	1387	EYS	2373
CNTN4	1361	CNTN5	2333

- * Principal Component AIMS
- * 0-5 and 95-100 percentiles of loadings distribution
- * ~33.000 SNPs for each component



Gene Family	MASSA	TOPPGENE
IL	0.000000000356	0.0001732
CYP	0.000000000592	-
CASR	0.000012504784	-
OPR	0.000031203482	-
ABCA	0.000247300015	-
APOLIPO	0.000473385685	-
NR	0.002542971	-
CALCR	0.004367941	-
MYOV	0.005819754	-
ALOX	0.010162726	0.0097402
ADH	0.013047684	0.017791
CASP	0.014487064	0.017791
CD, VEST, IGD	0.014487064	-
FANC	0.020223997	0.017791
SDRC2	0.028767948	-
RGS	0.031599638	0.0337145
CDK	0.037238712	-
CD, IGD	0.048420282	-
PRSS	0.053963137	-
SH2D	0.094535905	-
SLC	0.135954552	-
ANKRD	0.194559627	-
CD	-	0.0718073
ABC	-	0.0846113



* Next Steps

- * Improve annotation speed
 - * MySQL
 - * Java Otimization
- * Add new Databases
 - * SIFT, PolyPhen, Provean
 - * Epigenetic Databases
- * Integrate new tools and functionalities
 - * Networks