OASIS

A Novel Method of GWAS Analysis

Mohammad Saeed, MD

Search engine for Oasis in Gene Deserts

Saeed, M. Immunogenetics (2017) 69: 295-302. PMID: 28246883.

Python Code: https://github.com/dr-saeed/OASIS/blob/master/OASIS.py

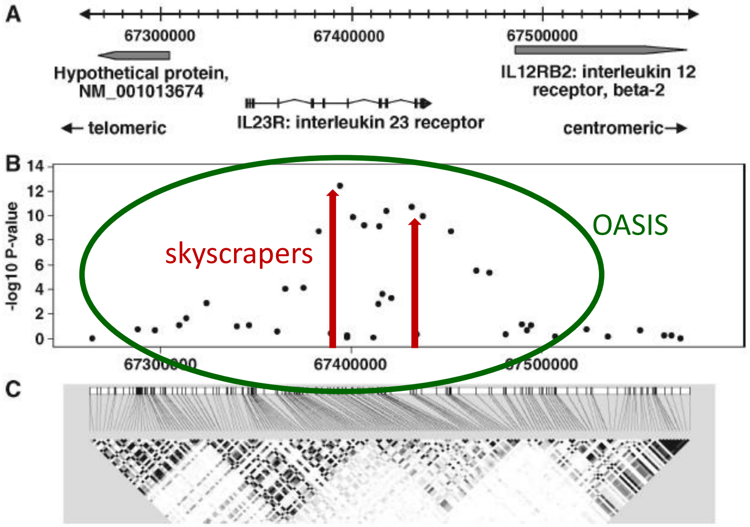


Figure modified from: Duerr et al. Science 2006, 314
M. Saeed. Immunogenetics (2017). PMID: 28246883.

Background

Complex Disease (e.g. SLE) susceptibility gene identification involves the following:

GWAS Principle: *complex disorders are caused by common variants (frequency >1%) and should therefore be detectable by linkage disequilibrium (LD) mapping using a large number of common variants.*

1. Low to modest effect sizes of susceptibility loci require large sample sizes. Hence small sample sized studies failed to detect SLE genes [PMID: 27933432]. *Replication and Joint analysis with multiple datasets became the norm in GWAS.*

2. Large sample sized studies were also plagued by phenotypic heterogeneity and statistical correction issues due to single variant analysis [PMID: 26502338, 27399966]. This has resulted in a few genes being consistently identified for complex disorders and a large number of candidate genes. Consequently, common variants explain about 15% of SLE heritability [PMID: 27933432, 26502338].

GWAS also led to the *problem of multiple testing* (PMID: 15272419, 28246883)

Two solutions were proposed:

- ***Statistical corrections***
- ***Gene-based testing***

3. ***Statistical corrections***:

Bonferroni, judged to be too stringent, led to the development of False Discovery Rate (FDR) [PMID: 12883005].

4. ***Gene-based Testing***:

To tackle the single variant statistical skewing (false positive associations) gene-based tests were proposed [PMID: 15272419, 20442747]. There are two major methods for gene-based testing

a) assigning the most significant p-value to a gene (e.g. MAGENTA. PMID: 20714348]
b) assigning a weighted p value to a gene [e.g. GATES/KGG. PMID: 21397060 , 21217833].

5. ***Locus-based testing:*** Enhancement of the Gene-based testing concept is locus-based testing incorporated in OASIS [PMID: 28246883, 29147756, 29923028].

Gene and locus based tests have the potential to detect susceptibility genes of modest effect sizes [PMID: 15272419].

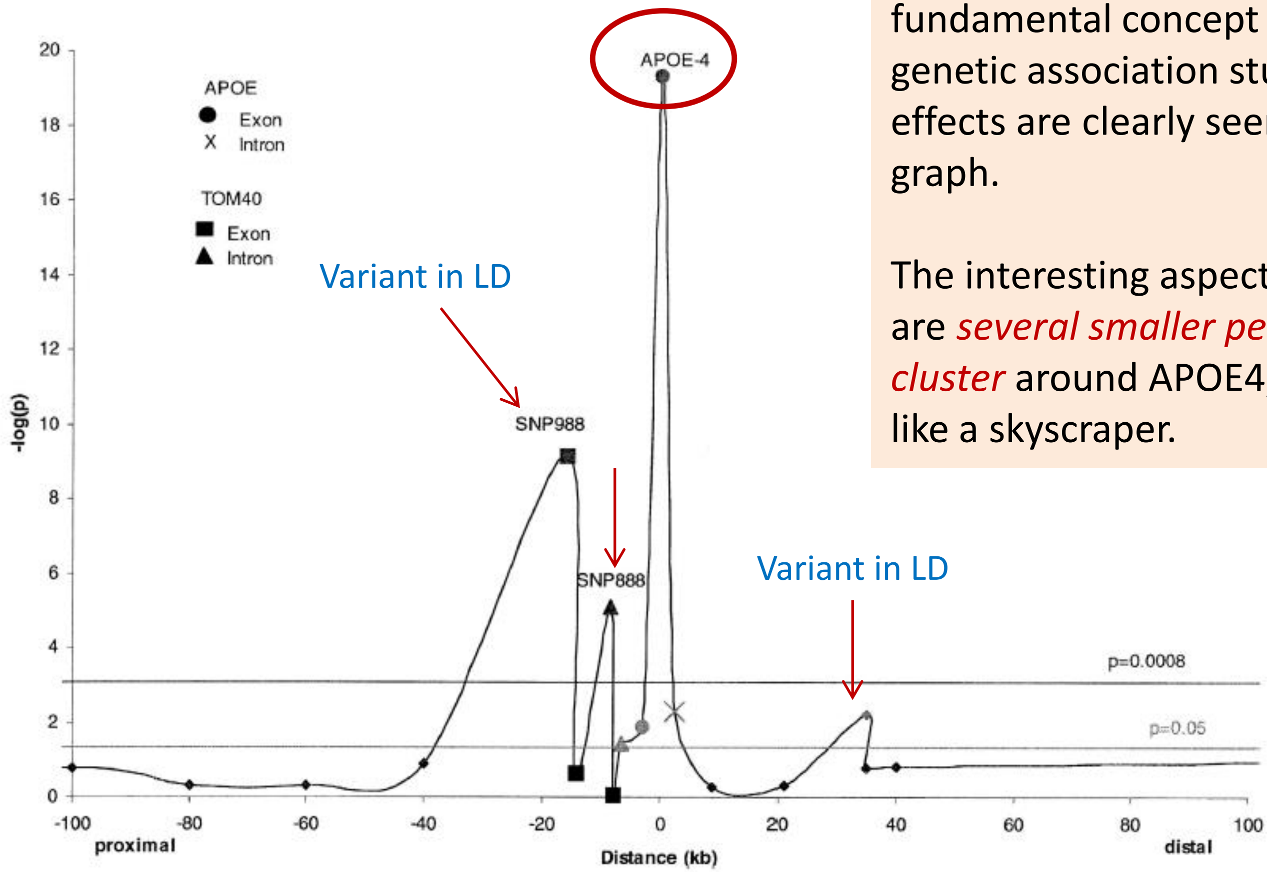
6. ***Genomic Convergence:***

This concept was introduced formally in Genomics in 2009 [PMID: 19241460] however, it is based on the simple scientific principle that multiple datasets with results pointing in the same direction is evidence of a true scientific finding. In Genomic studies the most frequent datasets that have been converged are genetic association studies highlighting a candidate gene and the expression of that gene in a biologically relevant tissue.

These principles are applied to identify novel complex disease (e.g. SLE) susceptibility genes

The OASIS concept

Martin et al.: Analysis of SNPs around *APOE*



True variant

Variant in LD

Variant in LD

Linkage disequilibrium (LD) is the fundamental concept driving genetic association studies. Its effects are clearly seen in this graph.

The interesting aspect is that there are *several smaller peaks* creating a *cluster* around APOE4, which stands like a skyscraper.

Figure 2 Plot of minus log of *P* value for case-control test for allelic association with AD, for SNPs immediately surrounding *APOE* (<100 kb).

Am. J. Hum. Genet. 67:383–394, 2000

Nature of Skyscrapers

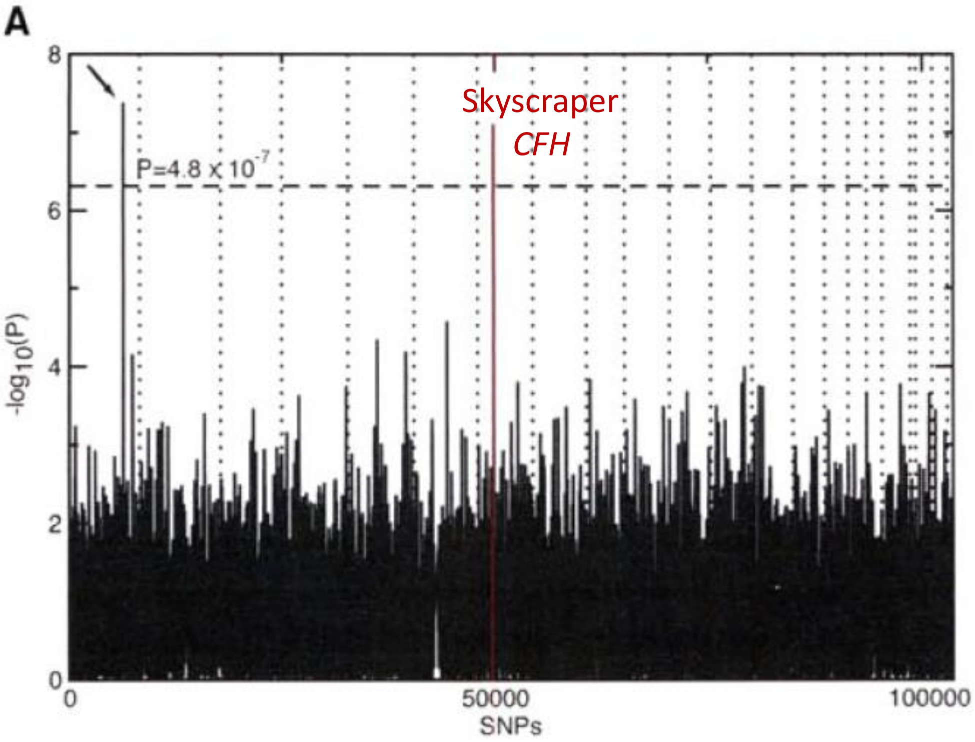
These are generated by pooling of mutants in the 'case' group (being absent in controls). The peaks are reduced by the merging of different mutations (carried on different haplotypes in the same gene).

Saeed et al. Neurology. 2009;72(19):1634-9. PMID: 19176896.

When complex phenotypes such as SLE are grouped together as 'cases' it is in fact a pooling of mutant haplotypes. It is nearly impossible to phenotypically accurately define mutant genotypes in such complex disorders.

As a result multiple haplotypes of varying intensity emerge. These may have the effect of canceling each other out based on the pattern of LD with the tested SNPs. This results in a modest peak, hard to confirm using stringent correction criteria such as Bonferroni. This has been the case in SLE GWAS as well [PMID: 28246883].

GWAS Manhattan plot



Most Genome-wide Association Studies (**GWAS**) are not as lucky as the CFH study which found a single extremely high association peak.

CFH data: Klein et al. Science 2005, 308

OASIS effect across GWAS

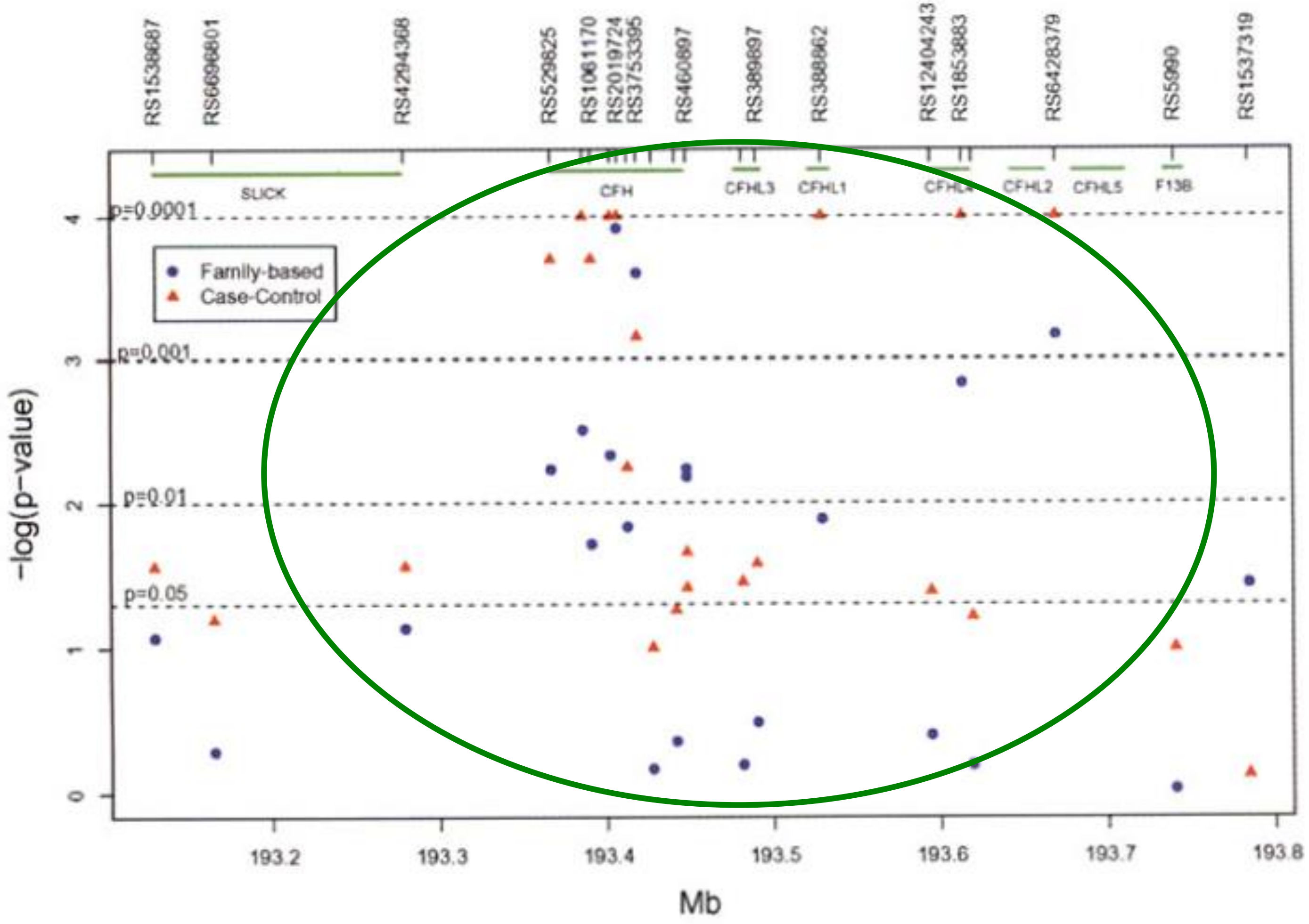


Fig. 2. Plot of family-based and case-control *P* values for all SNPs within the AMD-associated haplotype. The genomic region spanning each gene is indicated in green. $-\log_{10}$ of the nominal *P* values are plotted for each SNP. Results for both the family-based and case-control data sets converge within the *CFH* gene.

Similar observation in *CFH* data *(Haines et al. Science 2005, 308).*

This is due to LD with the functional variant.

I call this cluster the 'OASIS' in genomic 'deserts'.

OASIS: A Novel Method of GWAS Analysis

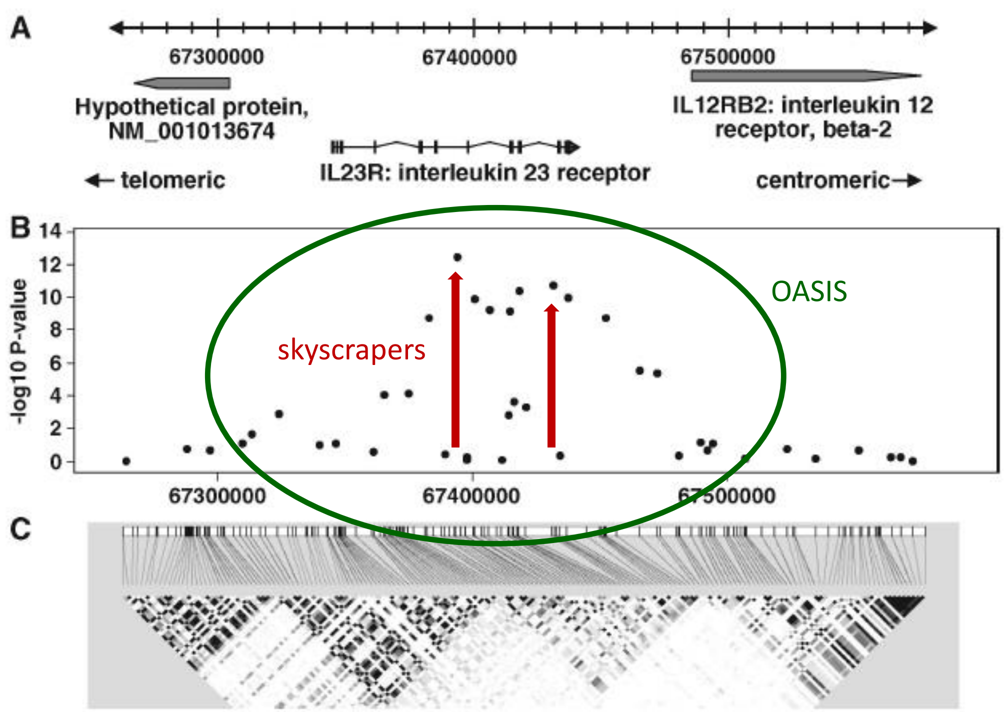


Fig. 1. Association signals in the *IL23R* gene region on chromosome 1p31. (**A**) Genomic locations of genes on chromosome 1p31 between 67,260,000 and 67,580,000 base pairs (Build 35). (**B**) The negative $\log_{10}$ association *P*-values (Cochran-Mantel-Haenszel chi-square test) from the combined Jewish and non-Jewish case-control cohorts are plotted for genotyped markers in the region. (**C**) Pairwise r^2 plot for International HapMap CEU data. The intensity of the shading is proportional to r^2. The *IL23R* gene is contained within two blocks of linkage disequilibrium, and the association signals are strongest in the centromeric block, which contains exons 5 to 11 and extends into the intergenic region between *IL23R* and *IL12RB2*. Note that markers in the block encompassing the *IL12RB2* gene do not demonstrate significant association.

The “OASIS” effect is clearly seen in the GWAS paper identifying *IL23R* as the gene for Crohn’s Disease *(Duerr et al. Science 2006, 314).*

Classically in GWAS, skyscrapers are searched for using varying cutoffs such as Bonferroni and FDR.

OASIS looks for LD clusters which are likely more robust evidence of the presence of a disease causing gene.

OASIS Python code:

https://github.com/dr-saeed/OASIS/blob/master/OASIS.py

Saeed, M. Immunogenetics (2017) 69: 295-302. PMID: 28246883.

OASIS vs Manhattan Plot: Similarities and Differences

OASIS

Structure of an 'OASIS' depends on:

- Linkage disequilibrium (LD) with the functional variant (LD may extend to 60kb)
- Number of SNPs genotyped in a genomic region (i.e larger OASIS with larger N of SNPs)
- Population stratification (between cases and controls)

Skyline (Standard association analysis)

Structure of Skyline (Manhattan plot) depends on:

- Allele frequency and sample size – information content of the allele (i.e low allele frequency variants can give high association signal due to the chi-squared distribution).
- LD with the functional variant (i.e high LD will give taller peaks).
- Population stratification

* Genotyping errors equally affect both methods and have to be corrected prior to any analysis.

If we use both methods in concordance with each other then greater information about disease causing genes can be extracted from the GWAS data.

How is OASIS analysis performed:

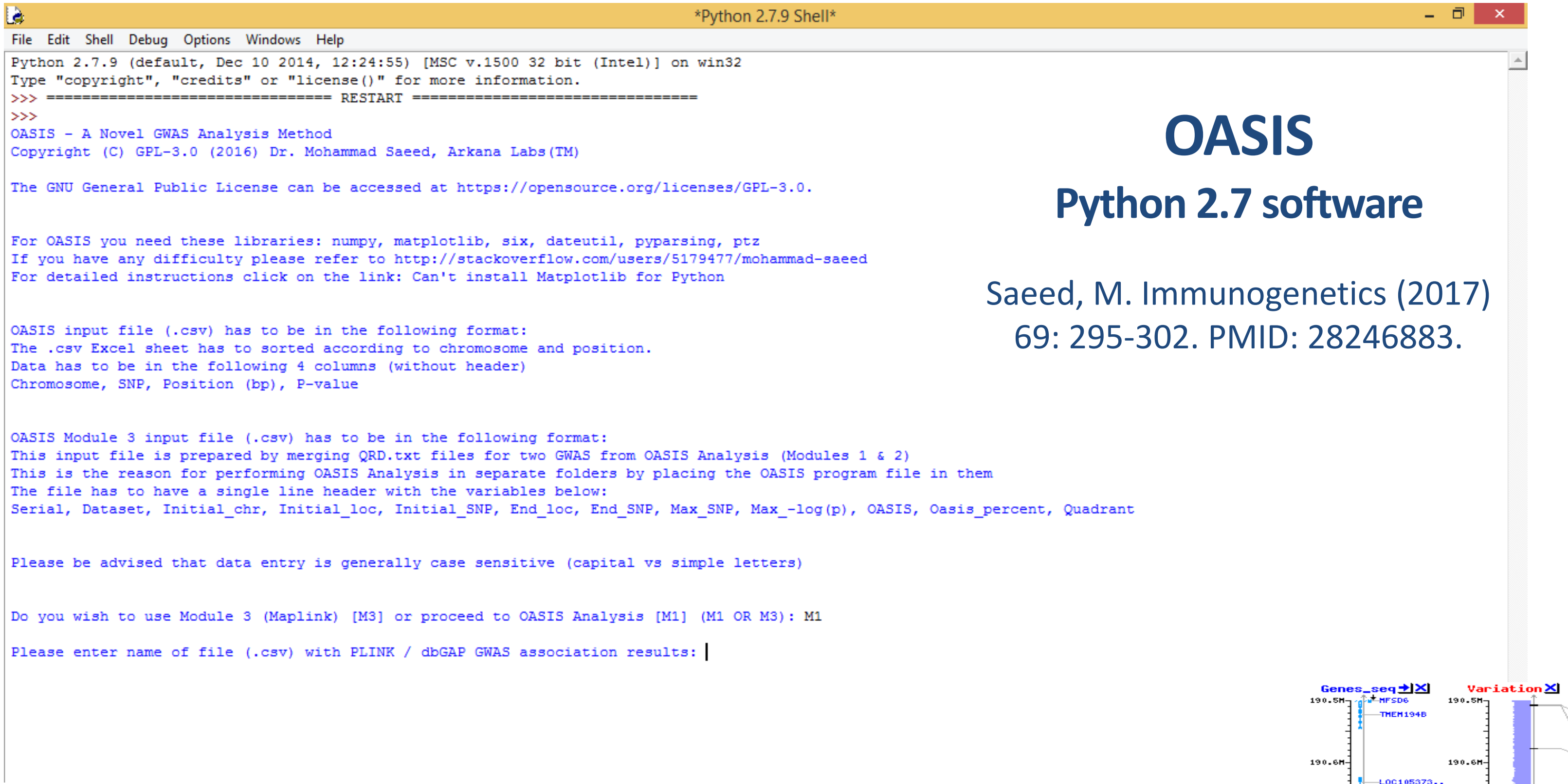
- The GWAS data is subjected to PLINK analysis to obtain P-values for all SNPs (Available through dbGAP in this format already).
- Window size is set at 200kb.
- The number of SNPs giving a P-value < 0.05 in this window are counted.
- This number forms the OASIS.
- The entire GWAS data is analyzed in this manner and two sets of data points are obtained in windows of 200kb: lowest P-values (plotted as –log P) and OASIS.
- The analysis can be performed for any window sizes as set by the user (200kb default)

Expected Results:

1. Plotting –log P values against OASIS scores generates Quadrants (next slide).
2. Regions with skyscrapers often have large OASIS around them. These can be mapped in concordance and represent the most significant disease associations (Quadrant A).
3. Some regions present as large OASIS but with no skyscrapers (Quadrant C).
4. Data in all three Quadrants (A, B, C) is significant based on 3-sigma cut off (as determined by the GWAS dataset).

- **Software:** The process has been automated for large GWAS datasets in my customized software (programmed in Python 2.7.9) and can be accessed below.

OASIS: https://github.com/dr-saeed/OASIS/blob/master/OASIS.py



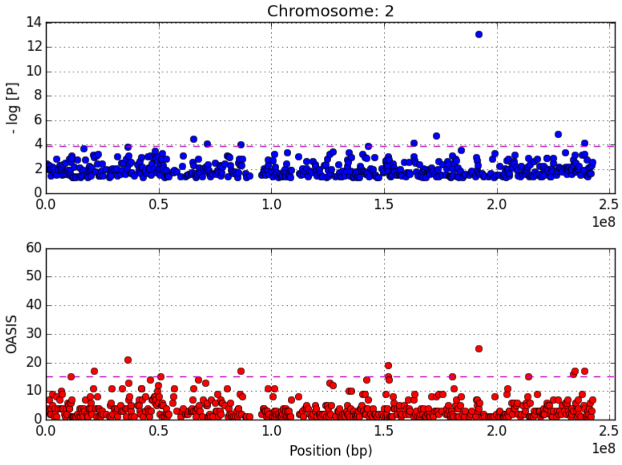
```
Python 2.7.9 (default, Dec 10 2014, 12:24:55) [MSC v.1500 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
OASIS - A Novel GWAS Analysis Method
Copyright (C) GPL-3.0 (2016) Dr. Mohammad Saeed, Arkana Labs(TM)

The GNU General Public License can be accessed at https://opensource.org/licenses/GPL-3.0.


For OASIS you need these libraries: numpy, matplotlib, six, dateutil, pyparsing, ptz
If you have any difficulty please refer to http://stackoverflow.com/users/5179477/mohammad-saeed
For detailed instructions click on the link: Can't install Matplotlib for Python


OASIS input file (.csv) has to be in the following format:
The .csv Excel sheet has to sorted according to chromosome and position.
Data has to be in the following 4 columns (without header)
Chromosome, SNP, Position (bp), P-value


OASIS Module 3 input file (.csv) has to be in the following format:
This input file is prepared by merging QRD.txt files for two GWAS from OASIS Analysis (Modules 1 & 2)
This is the reason for performing OASIS Analysis in separate folders by placing the OASIS program file in them
The file has to have a single line header with the variables below:
Serial, Dataset, Initial_chr, Initial_loc, Initial_SNP, End_loc, End_SNP, Max_SNP, Max_-log(p), OASIS, Oasis_percent, Quadrant


Please be advised that data entry is generally case sensitive (capital vs simple letters)


Do you wish to use Module 3 (Maplink) [M3] or proceed to OASIS Analysis [M1] (M1 OR M3): M1

Please enter name of file (.csv) with PLINK / dbGAP GWAS association results: 
```

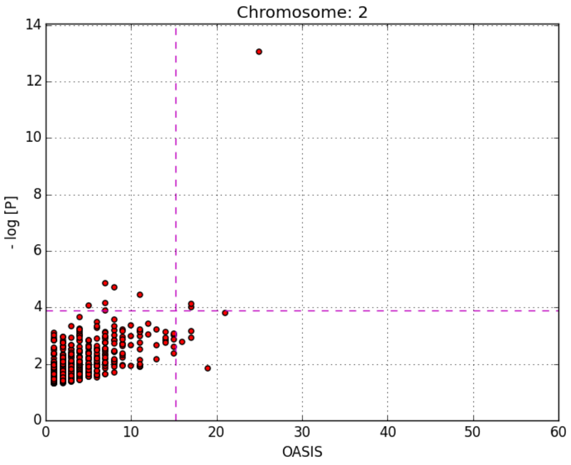
OASIS

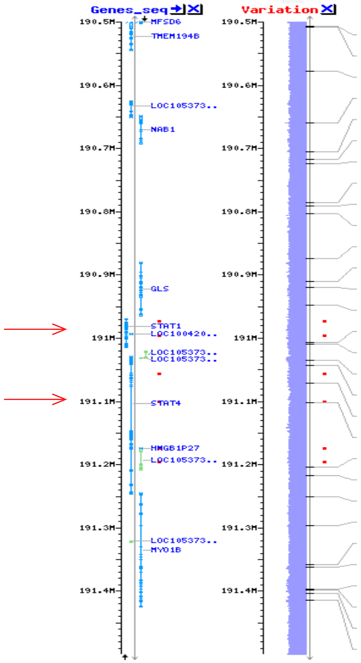
Python 2.7 software

Saeed, M. Immunogenetics (2017) 69: 295-302. PMID: 28246883.

Results e.g. Chromosome 2 – STAT1 / STAT4 for SLE

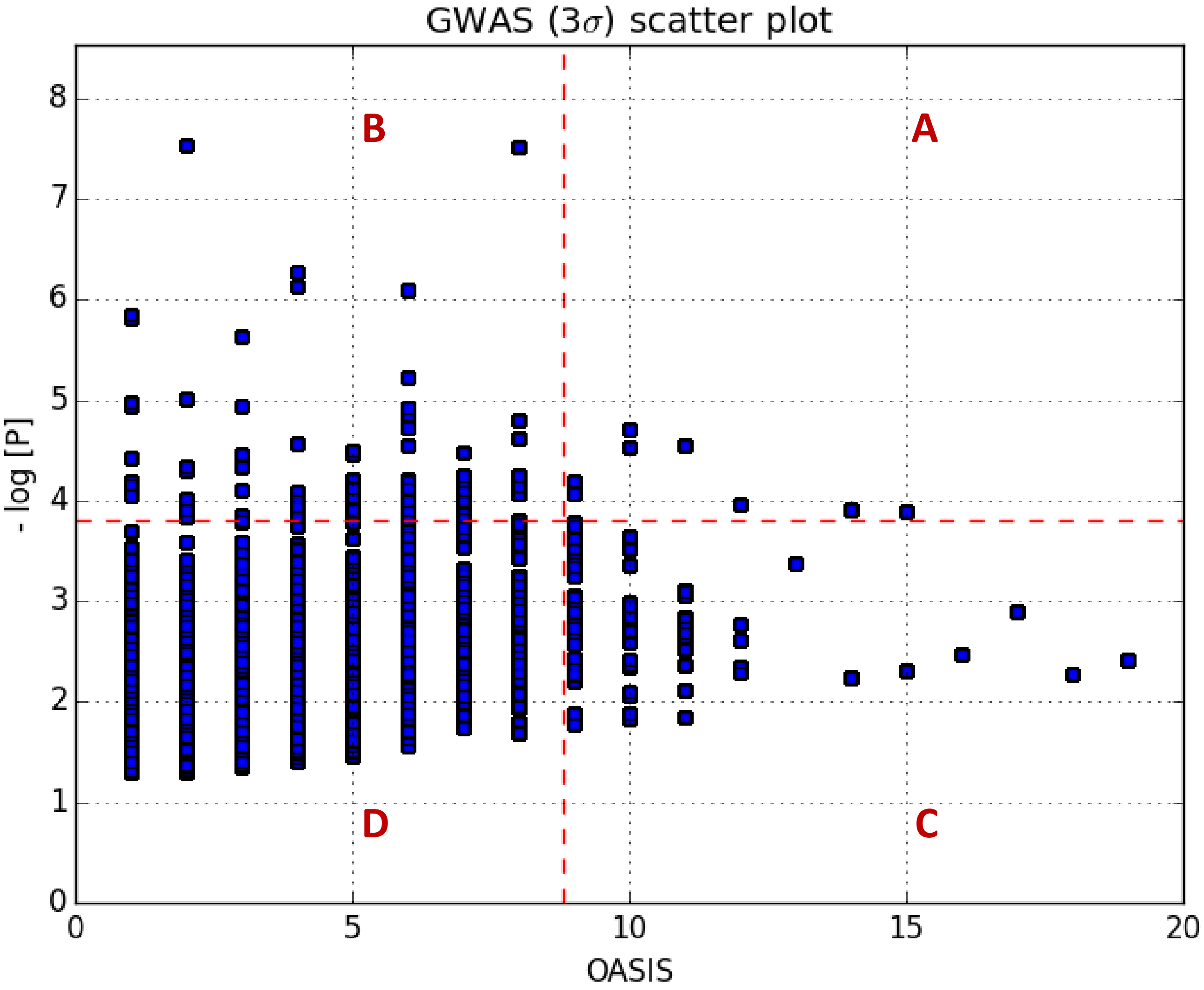






OASIS Quadrants Scatter plot

and applying the 3-sigma rule for determining significance



SLE GWAS OASIS Analysis.

Saeed, M. Immunogenetics (2017) 69: 295-302.

PMID: 28246883.

Quadrants

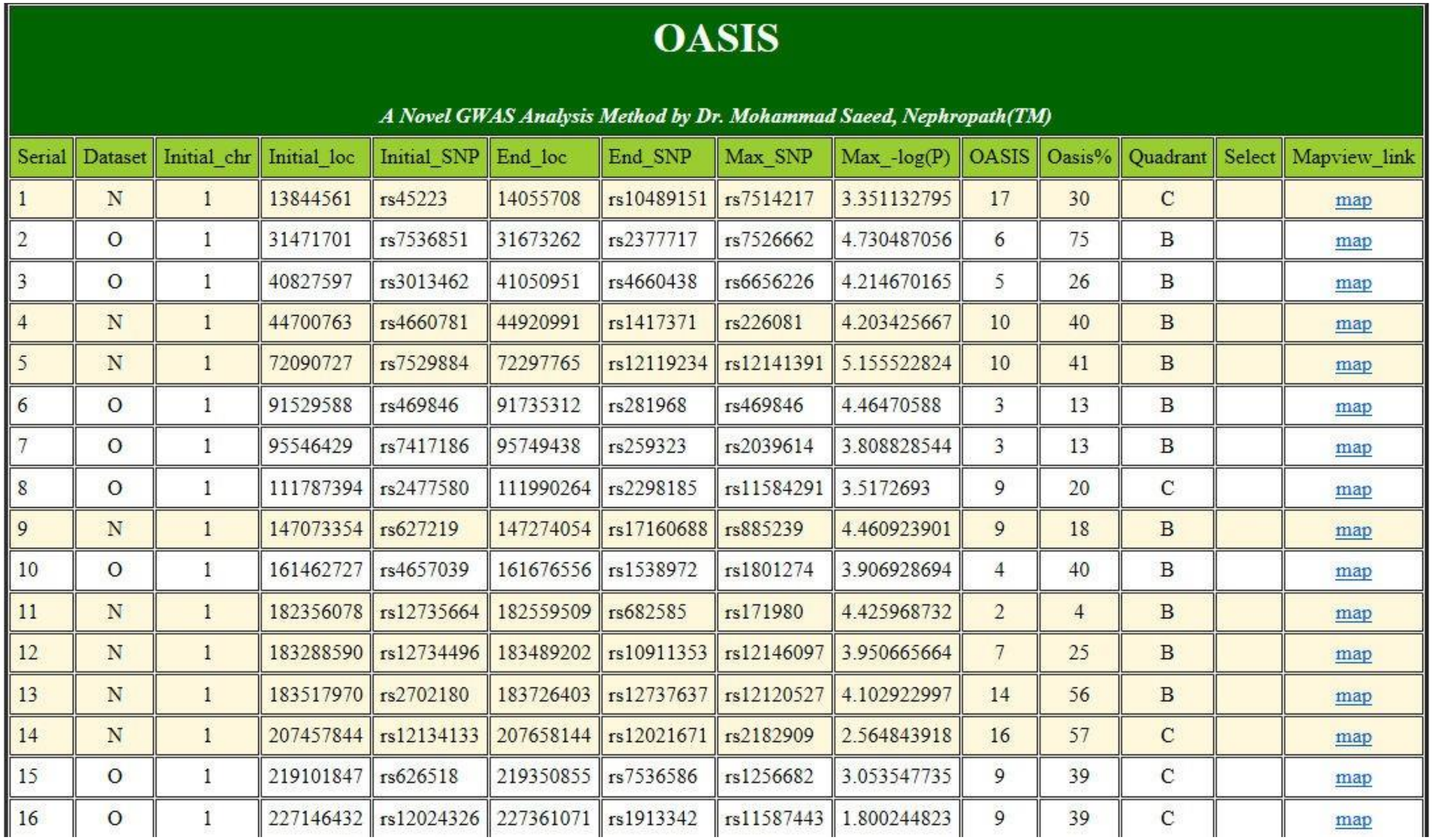
A: High oasis and –log[p]

B: High –log[p] only

C: High Oasis only

OASIS: Two GWAS Replication

Overlapping association blocks displayed as webpage with clickable Mapview links



OASIS

A Novel GWAS Analysis Method by Dr. Mohammad Saeed, Nephropath(TM)

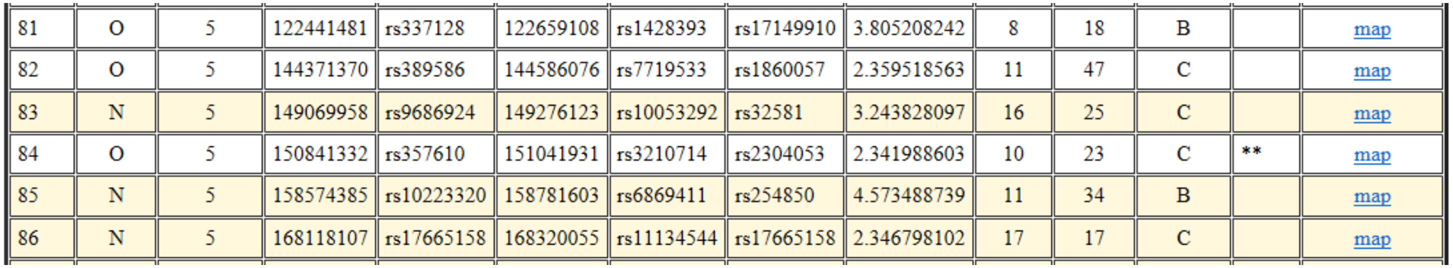
Serial	Dataset	Initial_chr	Initial_loc	Initial_SNP	End_loc	End_SNP	Max_SNP	Max_-log(P)	OASIS	Oasis%	Quadrant	Select	Mapview_link
1	N	1	13844561	rs45223	14055708	rs10489151	rs7514217	3.351132795	17	30	C		map
2	O	1	31471701	rs7536851	31673262	rs2377717	rs7526662	4.730487056	6	75	B		map
3	O	1	40827597	rs3013462	41050951	rs4660438	rs6656226	4.214670165	5	26	B		map
4	N	1	44700763	rs4660781	44920991	rs1417371	rs226081	4.203425667	10	40	B		map
5	N	1	72090727	rs7529884	72297765	rs12119234	rs12141391	5.155522824	10	41	B		map
6	O	1	91529588	rs469846	91735312	rs281968	rs469846	4.46470588	3	13	B		map
7	O	1	95546429	rs7417186	95749438	rs259323	rs2039614	3.808828544	3	13	B		map
8	O	1	111787394	rs2477580	111990264	rs2298185	rs11584291	3.5172693	9	20	C		map
9	N	1	147073354	rs627219	147274054	rs17160688	rs885239	4.460923901	9	18	B		map
10	O	1	161462727	rs4657039	161676556	rs1538972	rs1801274	3.906928694	4	40	B		map
11	N	1	182356078	rs12735664	182559509	rs682585	rs171980	4.425968732	2	4	B		map
12	N	1	183288590	rs12734496	183489202	rs10911353	rs12146097	3.950665664	7	25	B		map
13	N	1	183517970	rs2702180	183726403	rs12737637	rs12120527	4.102922997	14	56	B		map
14	N	1	207457844	rs12134133	207658144	rs12021671	rs2182909	2.564843918	16	57	C		map
15	O	1	219101847	rs626518	219350855	rs7536586	rs1256682	3.053547735	9	39	C		map
16	O	1	227146432	rs12024326	227361071	rs1913342	rs11587443	1.800244823	9	39	C		map

Key:

- Composite data from two GWAS datasets
- If the ‘map’ hyperlink is right clicked (open in new tab) it will open the location of the region in Mapview with the SNPs highlighted
- Max –log(P) is the ‘Max_SNP’ P-value as –log
- OASIS – the number of SNPs in a 200kb block that had $P<0.05$
- OASIS % is the percentage of significant SNPs (OASIS) relative to the number of SNPs genotyped

OASIS Replication

Overlapping Regions within 2Mb (i.e. close by) in both GWAS datasets



81	O	5	122441481	rs337128	122659108	rs1428393	rs17149910	3.805208242	8	18	B		map
82	O	5	144371370	rs389586	144586076	rs7719533	rs1860057	2.359518563	11	47	C		map
83	N	5	149069958	rs9686924	149276123	rs10053292	rs32581	3.243828097	16	25	C		map
84	O	5	150841332	rs357610	151041931	rs3210714	rs2304053	2.341988603	10	23	C	**	map
85	N	5	158574385	rs10223320	158781603	rs6869411	rs254850	4.573488739	11	34	B		map
86	N	5	168118107	rs17665158	168320055	rs11134544	rs17665158	2.346798102	17	17	C		map

Saeed, M. Immunogenetics (2017)
69: 295-302. PMID: 28246883.

Key:

- If an OASIS hit (in the html table – previous slide) in one GWAS dataset (e.g 'O') is close (less than 2Mb) to another hit in the second GWAS dataset (e.g. 'N) then it is marked with ** in the select column In the example above serial numbers 83 and 84 are 'selected'.
- The ** will always be on the lower hit i.e it will be less than 2Mb away from the hit above it