MISSION-T2D

Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data

Work Package 3

Deliverable 3.2

Applying the approach to patient network at molecular and whole body

Document Information

Grant Agreement	**Nº**	600803	**Acronym**	MISSION-T2D
Full Title	Multiscale Immune System SImulator for the Onset of Type 2 Diabetes integrating genetic, metabolic and nutritional data			
Project URL	http://www.mission-t2d.eu			
EU Project Officer	**Name**	Dr. Adina Ratoi		

Deliverable	**No**	3.2	**Title**	Analysis of gene copy number, SNPs and other omics
Work package	**No**	3	**Title**	

Date of delivery	**Contractual**		**30.08.2014**		**Actual**		**02.09.2014**		
Status	**Version 1.2**				**Final**		**1.2**		
Nature	**Prototype**		**Report**	**X**	**Dissemination**		**Other**		

Dissemination level	Consortium+EU	
	Public	X

Target Group (If Public)		Society (in general)	
Specialized research communities	X	Health care enterprises	
Health care professionals		Citizens and Public Authorities	

Responsible Author	**Name**	Pietro Lio	**Partner**	UniCAM
	Email	pl219@cam.ac.uk		

Version Log

Issue Date	Version	Author (Name)	Partner
20.08.2014	1.0	Pietro Lió	UniCAM
25.08.2014	1.1	Filippo Castiglione, Paolo Tieri	CNR
02.09.2014	1.2	Pietro Liò	UniCAM

Executive Summary	The analysis and the integration of omics data (gene copy number, SNPs, gene expression, pathways, methylation) related to diabetes and inflammatory diseases has required building a pipeline of novel programs and develop new algorithms that exploit the gene family structure of identified over expressed genes. Particularly relevant are the comparative omic analysis for the chemokine receptors and ligands. Two papers are in completion.
Keywords	Multi Omic analysis, CNV, phylogeny, chemokines

Contents

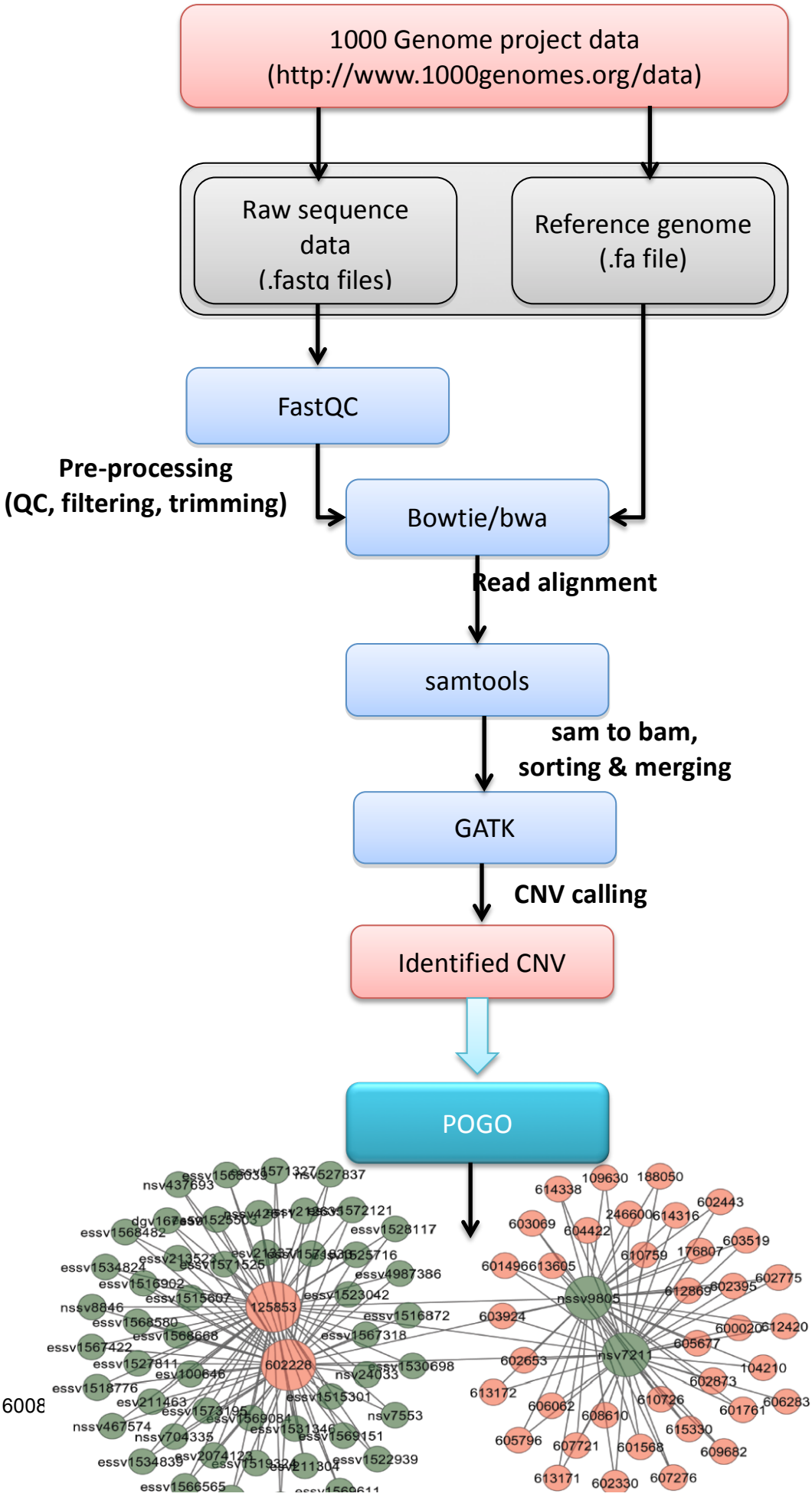
1 Introduction

Diseases with similar genetic, environmental, and lifestyle risk factors may be co-morbid in patients or may be risk factors for additional conditions [6]. Shared risk and environmental factors have similar consequences, prompting the co-occurrence of related diseases in the same patient. For an instance, many well-known and influential environmental factors such as smoking, diet, and alcohol intake are strongly associated with diabetes type 1 and type 2, and obesity. Also, many serious chronic diseases, such as cancer and diabetes, are complex diseases influenced by a combination of environment and epistasis between many genes. Therefore, a patient diagnosed for a combination of diseases and exposed to specific environmental, lifestyle and genetic risk factors may be at a considerable risk of developing several other genetically and environmentally related diseases. It is now well accepted that phenotypes are determined by genetic material under environmental influences. Increasing evidence has revealed that microRNAs (miRNAs) play important roles in the development and progression of human diseases. Functionally related miRNAs tend to be associated with phenotypically similar diseases [10]. Recently, genome-wide association studies (gwas) have proved useful as a method for exploring phenotypic associations with diseases. Single-nucleotide polymorphisms (SNPs), a variation of a single nucleotide, are assumed to play a major role in causing phenotypic differences between individuals. It has become possible to assess systematically the contribution of common SNPs to complex disease. In copy number variations (CNVs) longer stretches of DNA can get lost, duplicated, or rearranged in the genome of an individual that cause various phenotypic abnormalities. CNVs are significantly associated with the risk of complex human diseases including inflammatory autoimmune disorders, diabetes etc. The development of type 2 diabetes has also been known to be influenced by genetic and environmental factors. In this way, diseases may share many distinct types of relationships with varying levels of risk for disease comorbidity. Thus, a singular view of dependencies among diseases is not sufficient. As more and more ontology, phenotype, omics and environmental data sets become publicly available, it is beneficial to improve our understanding of human diseases and diseases comorbidities based on these new system-level biological data. The integration analysis of various 'omic' data has become increasingly widespread because each approach has intrinsic caveats. For instance, important information may be missing because of false

negatives or misleading because of false positives. Some studies have indicated that these limitations can be mitigated by integrating two or more omic datasets. Several studies have reported on the role of a single molecular or phenotypic measure to capture disease-disease relationships (such as shared genes or gene ontology), but a comprehensive understanding requires to inspect multiple sources of evidence including miRNA-based relationships, shared environmental factors, ontology, SNPs, CNVs and phenotypic manifestations.

2 Copy number variation

Whole genome sequencing enables a high resolution view of the human genome and provides unique insights into genome structure at an unprecedented scale. Genetic variation in the human genome occurs in many forms ranging from large chromosomal abnormalities to single nucleotide variations, each with varying functional significance. DNA copy number variation (CNV) has been recognised as an important source of genetic variation. Copy number variation (CNV) is one such genetic variation that can range from a few kilobases to megabases, and involve deletions, duplications, insertions or translocations. The associations between CNVs and phenotypic variation or disease-susceptibility are increasingly being investigated, with the most obvious mechanism being gene-dosage caused by variations in the number of copies of a gene or its associated regulatory elements. Several methods and tools can be used for determining CNV information based on sequencing data. To characterise the landscape of structure variations, we used whole genome sequencing data from publicly available 1000 genomes project (http://www.1000genomes.org/data#DataAccess). The workflow of the CNV sequence data from the 1000 genomes database is shown in figure 1. There have been a number of tools to infer copy number variation in the genome from the sequence data. In our pipeline we have used different unix based software tools for the analysis of the data to identify the CNV for the type 2 diabetes. We have used FastQC tool for the qualitity control, filtering and trimming of the raw sequence data. Bowtie is used for mapping of type 2 diabetes sequence with the reference sequence, and samtools is used for converting sam to bam format of the data, and sorting and merging of the reads. Finally, GATK toolkit is used for CNV calling to identify the CNV for the type 2 diabetes. Then we used the result of the CNV analysis for type 2 diabetes as input of the procedure to identify the inflammatory links for the type 2 diabetes.


1000 Genome project data
(http://www.1000genomes.org/data)
Raw sequence data
(.fastq files)
Reference genome
(.fa file)
FastQC
Pre-processing
(QC, filtering, trimming)
Bowtie/bwa
Read alignment
samtools
sam to bam,
sorting & merging
GATK
CNV calling
Identified CNV
POGO

The figure above shows the pipeline we have built to analyse CVN in multi omic data. This represents a substantial improvement with respect to previous research (Bae, 2011; see also McCarroll, 2007). The final graph is the output of the pipeline focusing on "Type 2 Diabetes mellitus" (Omim identification is 602228), which is used as input to the comorbidity CNV with inflammatory and other diseases (paper in progress). The introduction of the biomedical informatics library tools (Gene ontology and text mining) provides means to extract data from medical literature and more output.

```
> comorbidityCNV("602228", "OMIM")

SYMBOL    OMIM ENTREZID  PATH    GO            EVIDENCE
TCF7L2  602228    6934    04310  GO:0005515         IPI
TCF7L2  602228    6934    04310  GO:0005515         IPI
TCF7L2  602228    6934    04310  GO:0005515         IPI
TCF7L2  602228    6934    04310  GO:0005515         IPI
TCF7L2  602228    6934    04310  GO:0005515         IPI
 ...   ...

ONTOLOGY   CNV.ID  Chr    Start       End       VarSubtype
MF         nsv7211   10  108617417  118351740  Inversion
MF         nsv7553   10  114845707  114890646  Loss
MF      esv2074123   10  114876971  114877374  Deletion
MF        nsv24033   10  114877162  114877217  Loss
MF       nsv527837   10  114888608  114911079  Loss
 ...   ...
```

The figure above shows an example of the output statistics of comorbidity(CNV("602228" , "OMIM")). The OMIM disease id of the "Type 2 Diabetes mellitus" is 602228, which is used as input to the comorbidity CNV. We show disease comorbidity for the "Type 2 Diabetes mellitus" through the CNVs-disease associations. Then we have considered different datasets of inflammatory, T2D diabetes and T1D diabetes and we have computed the over expressed and under expressed genes. The datasets used are: Rheumatoid arthritis (GSE1919); Osteoarthritis condition (GSE1919); T2D(GSE9006); T1D(GSE9006). Normalization procedures and statistical analysis are performed by using Bioconductor R packages (Gentleman, 2004); the background correction and normalization is performed by using (Therneau et al, 2007) algorithm. PLIER algorithm produces an improved gene expression value as compared to the other algorithms. It accomplishes this by incorporating experimental observations of

feature behavior. Specifically, it uses a probe affinity parameter, which represents the strength of a signal produced at a specific concentration for a given probe. The probe affinities are calculated using data across arrays. The Bioconductor package Limma was also used to calculate average expression levels, log fold changes and adjusted p-values for each probe. Standard anova and Box plots representation were used to analyze and check out visually the expression levels of these genes for different conditions. This below is a table (2.1) showing the overexpressed and underexpressed genes in the four datasets.

We found that NFKB, Signalling pathways, tgfbeta, TNF and chemokine family of genes appear to be common among these diseases (paper submitted). The gene families will be analysed using novel methods, explained in the next section.

T1D (PBMCs)		T2D (PBMCs)		Osteoarthritis		Rheumatoid arthritis	
Gene Symbol	P.Value	Gene Symbol	P.Value	Gene Symbol	P.Value	Gene Symbol	P.Value
TNFAIP3	1.77E-07	TRIM2	4.75E-05	NFKBIA	5.30E-05	TGFBR3	2.40E-08
NFKBIA	1.37E-05	AKT2	0.000173708	TNFAIP3	8.45E-05	MAPK13	6.35E-06
TNFSF14	0.000356257	AKT3	0.000222778	TGFBR3	0.000120087	TGFA	7.71E-06
TNFRSF17	0.000850768	RELA	0.000299556	MAPK13	0.001040518	TRIM2	0.000182868
CCR5	0.00089995	TNFAIP1	0.00063572	TGFA	0.002815045	TNFSF9	0.000363318
TGFBR2	0.001134511	NFKBIA	0.000791186	TNFSF9	0.003727685	TNFRSF1B	0.000479444
PDGFC	0.004640007	TGFB1	0.000835585	BMP2	0.004106045	TNFRSF17	0.000514263
TNFRSF1A	0.004647921	MAP3K7	0.00084375	IFI27	0.005217842	TNFAIP6	0.000542038
MAPK13	0.007592876	REL	0.000888489	TNFRSF11A	0.008025001	IFI27	0.000577389
TNFAIP1	0.009489591	TNFRSF14	0.001387678	AKT3	0.010452435	CCR5	0.000586357
NFKB2	0.009642252	TNFSF14	0.001827428	TNF	0.012861793	SQSTM1	0.0016662
NUP62	0.010822798	TRAF6	0.002180766	TGFB2	0.015822861	RELB	0.001691721
TNFRSF13B	0.013286132	TNFAIP3	0.002634127	TGFBR1	0.023892893	TNFSF10	0.001724084
TNFSF18	0.015538582	TGFA	0.002730714	RELA	0.026085929	TGFBI	0.001929109
TNFRSF10C	0.016132998	MAPK14	0.003496727	TGFB3	0.030358653	MAPK10	0.001939754
AKT3	0.018507548	NFKB2	0.003812117	IL6	0.034326865	TNFSF11	0.001986958
TGFBI	0.018586844	NFKB1	0.003953564	TGFBR2	0.037806678	TNF	0.002813131
MAPK4	0.030868863	TNFSF10	0.004168785	IGF1R	0.038400057	MAPK1	0.002855606
CSF1	0.03300604	TNFAIP2	0.010169047	TNFRSF1B	0.04339234	NFKB1	0.003135214
TNFAIP6	0.033038618	CSF1	0.013497857	PDGFRA	0.044040037	LRP5	0.00361201
TRAF6	0.034594707	MAPK1	0.015620066	TNFSF11	0.04904459	TGFBR1	0.007130922
BMPR2	0.035076531	NFKB2	0.019309942			TNFSF8	0.016486512
LRP6	0.037451428	TNFRSF13B	0.020690122			MAPK7	0.017046898
SQSTM1	0.037820278	PDGFRA	0.022393241			TNFRSF25	0.020331139
TGFB3	0.041756839	MAPK12	0.02749298			TNFSF14	0.024364697
PDGFA	0.042711089	MAPK3	0.028945413			NUP62	0.027359276
TGFA	0.046157749	AKT1	0.029425624			TNFAIP8	0.030460145
IFI44L	0.047610377	MAPK11	0.031190453			CSF1	0.032119483
		TNFRSF8	0.031377416			TNFRSF11B	0.032600528
		TNFRSF10B	0.032735653			TNFRSF9	0.033094969
		IL6	0.036438795			NFKBIA	0.047416796
		TNFRSF21	0.038752772				
		TGFBR2	0.046911857				
		IFI44L	0.047309419				

Table 2.1

3 Pathways, methylation, gene expression data integration

The previous analysis has revealed that the most overexpressed genes belong to NFKB, Signalling pathways, tgfbeta, TNF and chemokine gene family. The sequence similarities within each gene family provides a ground for applying phylogenetic

methods. We have analysed different families of genes involved in diabetes and inflammatory processes. Here we show for the case of the chemokine family. Felsenstein proposed the first continuous phylogenetic comparative methods procedure, independent contrasts. This assumes a Brownian motion evolving on the phylogenetic tree and we observe the tree and contemporary phenotypic sample. It was observed that such a model does not allow for modeling adapting traits. A Brownian motion will mean that the phenotype randomly oscillates around the ancestral state. Therefore an Ornstein–Uhlenbeck process was proposed with different regimes on the phylogeny to model a phenotype adapting to different conditions (e.g. habitats). This was further developed to a trait evolving towards a randomly evolving environment. Ornstein–Uhlenbeck models have been also applied to study evolutionary rates .

One can naturally take the evolving phenotype to be measured gene–expression levels and apply the aforementioned levels. However it would be more interesting to consider how expression levels of different genes co–evolve. In an Ornstein–Uhlenbeck model for a multiple, say k, co-adapting traits is presented, $dY(t) = -A(Y(t) - \theta(t))\,dt + \Sigma dB(t)$, where A, Σ are k×k matrices, θ is a vector step function over the phylogeny and B(t) is a k–dimensional standard Wiener process. The maximum–likelihood estimation procedure can further be combined into the estimation procedure measurement error (or intra–species variability). This is an important factor to keep in mind as micro–array experiments can be very noisy and measurement variance can have a profound effect on a phylogenetic analysis (paper in progress; see also Bartoszek et al, 2011; Butler et al, 2004). The data we have analysed is the chemokine receptors and chemokine ligands. 23 chemokine receptors were collected from NCBI Gene database manually CCBP2, CCR1, CCR10, CCR2, CCR3, CCR4, CCR5, CCR6, CCR7, CCR8, CCR9, CCRL1, CCRL2, CMKLR1, CX3CR1, CXCR1, CXCR2, CXCR3, CXCR4, CXCR5, CXCR6, CXCR7, XCR1. 46 chemokine ligands and their binding information with the corresponding receptors were obtained from (reference). The binding information between the ligands and receptors are shown in the Table 3.1.

Table 3.1 Chemokine receptors and the binding ligands

Receptor	Binding chemokine ligands
CX3CR	CCL26;CX3CL11
XCR1	XCL1;XCL2
CXCR6	CXCL16
CXCR5	CXCL13

CXCR4	CXCL12
CXCR2	CXCL1;CXCL2;CXCL3;CXCL5;CXCL6;CXCL7;CXCL8
CXCR1	CXCL6;CXCL7;CXCL8
CCR9	CCL25
CCR8	CCL1;CCL16;CCL18
CCR6	CCL20
CCR4	CCL17;CCL22
CCR3	CCL3L1;CCL5;CCL7;CCL11;CCL13;CCL15;CCL24;CCL26;CCL28;CXCL9;CXCL10;CXCL11
CCR10	CCL27;CCL28
CCR7	CCL19;CCL21
CXCR3	CCL11;CXCL9;CXCL10;CXCL11;CXCL4
CCR2	CCL2;CCL7;CCL8;CCL11;CCL13;CCL16;CCL24;CCL26
CCR5	CCL3;CCL3L1;CCL4;CCL4L1;CCL5;CCL7;CCL8;CCL11;CCL13;CCL14;CCL16;CCL26;CXCL11
CCR1	CCL3;CCL3L1;CCL4;CCL5;CCL7;CCL8;CCL13;CCL14;CCL15;CCL16;CCL23;CCL26;CCL6;CCL9;CCL10
CXCR7	CXCL11;CXCL12

For all the members we have considered the gene expression and methylation data for diabetes and inflammatory diseases. We summarise the following data collection and results:

a) Gene expression in diabetes

GSE9006: Gene expression in peripheral blood mononuclear cells (PBMCs) from children with diabetes measured by Affymetrix HGU133A, including 24 healthy samples (Health), 43 type 1 diabetes patients (T1D) and 12 type 2 diabetes patients (T2D).
The raw data was downloaded from GEO database and was processed by using RMA method in the “affy” package from Bioconductor.

b) DNA methylation in diabetes

GSE34008: DNA methylation profiling of whole blood were measured by using Illumina's Infinium HumanMethylation27 Beadchip array. The dataset encompasses profiles of 12 non-diabetic control blood donors and 12 type-2 diabetic (T2D) individuals. GSE56606: Genome-wide DNA methylation profiles of purified CD14 and CD4 monocytes were generated by using HumanMethylation27 Beadchip array from monozygotic (MZ) twin pairs (50% T1D onset pairs and normal pairs). There are 100 samples in total, including 17 T1D samples and 35 normal samples in CD14, and 15 T1D samples and 33 normal samples in CD4. The methylation data were extracted from Series Matrix Files which are downloaded from GEO database. The methylation

levels (beta values) are the ratios of methylated signals (M) with the corresponding total signals (M+U).

Both the gene expression and methylation data of diabetes were downloaded from TCGA.

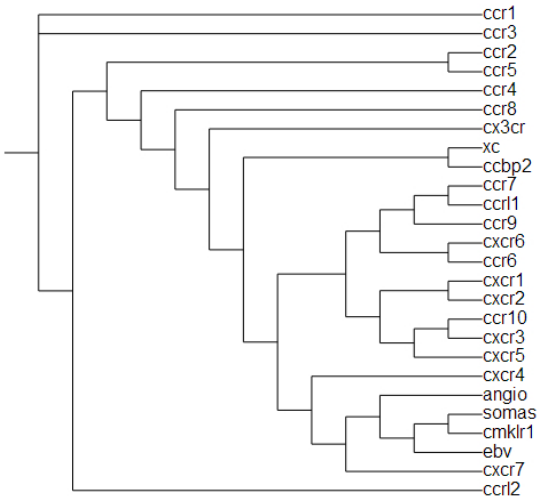


Figure 3.1 Phylogenetic tree using a Bayesian approach (other approaches have been investigated)

Differential expression and methylation

Differentially expressed genes (DEG) and differentially methylated genes (DMG) were identified by using T test, with p-value<0.05. (Because of the weak significance, there is not DMG after the BH correction. Here, the DEG and DMG results are not corrected by FDR correction.) As shown in Table 3.1, “0” represents that the preceptor is not differentially expressed/methylated, and “1” means the opposite. CCR4, CCRL1, CCRL2 and CXCR2 are all differentially expressed in both T1D and T2D. CCR2, CCR3, CCR5 and CX3CR1 only differentially expressed in T2D, while CXCR3 only differentially expressed in T1D. There are only 3 receptors that are differentially methylated in T1D or T2D, and there is not overlap between DEGs and DMGs, which may suggest that the chemokine receptors have different expression and methylation patterns during the genesis of diabetes.

Table 3.2 Differentially expressed/methylated receptors in diabetes

Receptor	Expression		Methylation		
	T1D vs N	T2D vs N	T1D vs N cd4	T2D vs N cd14	T2D vs N

CCBP2	0	0	0	0	0
CCR1	0	0	0	0	0
CCR10	0	0	0	1	0
CCR2	0	1	NA	NA	NA
CCR3	0	1	0	0	0
CCR4	1	1	0	0	0
CCR5	0	1	NA	NA	NA
CCR6	0	0	0	0	0
CCR7	1	0	0	0	0
CCR8	0	0	0	0	0
CCR9	0	0	0	0	0
CCRL1	1	1	NA	NA	NA
CCRL2	1	1	0	0	0
CMKLR1	0	0	0	0	0
CX3CR1	0	1	0	0	0
CXCR1	0	0	0	0	0
CXCR2	1	1	0	0	0
CXCR3	1	0	0	0	0
CXCR4	0	0	0	0	0
CXCR5	0	0	0	1	0
CXCR6	0	0	0	0	1
CXCR7	0	0	0	0	0
XCR1	0	0	0	0	0

Expression and methylation patterns of chemokine receptors

The heatmaps show the expression and methylation of the receptors in diabetes. The dark color (red) means the low expression/methylation level, and the lighter color means the higher level. From these figures, we will get some results as follows:

a) There is big inconsistence between expression and methylation of the receptors.
b) The methylation patterns in CD4 cells and CD14 cells of T1D are not the same, but the clusters of the receptors are similar.
c) There are slight differences in the methylation patterns between T2D and T1D.

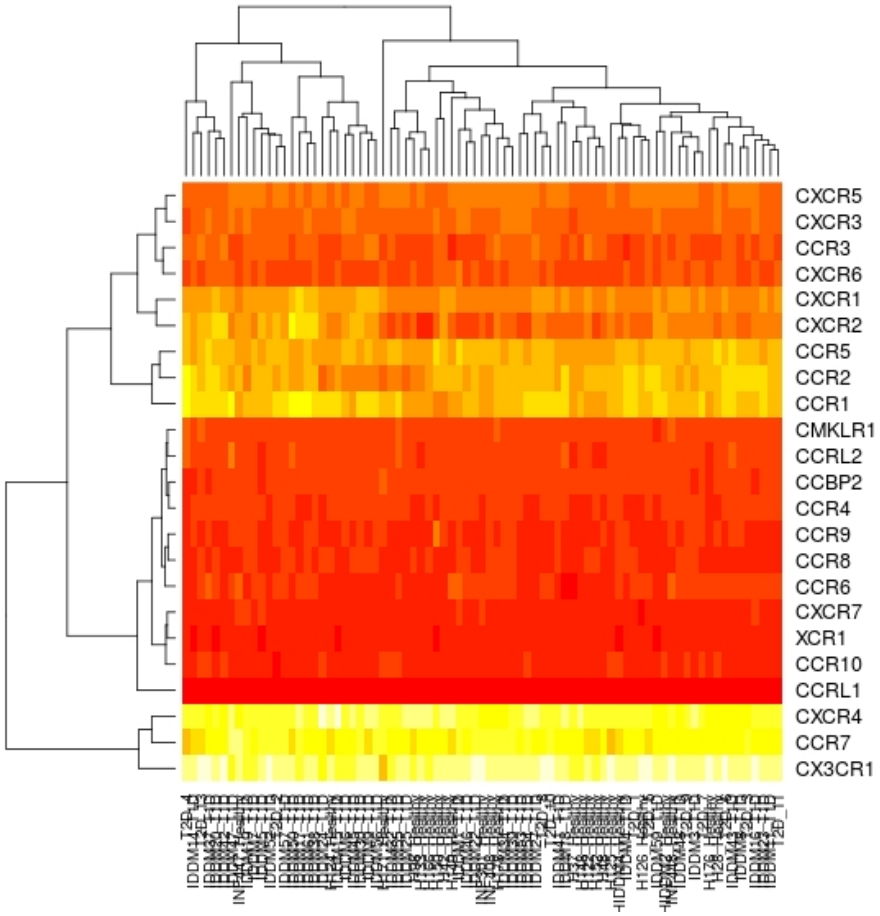


Figure 3.2 Heatmap of the expression of receptors in Normal, T1D and T2D

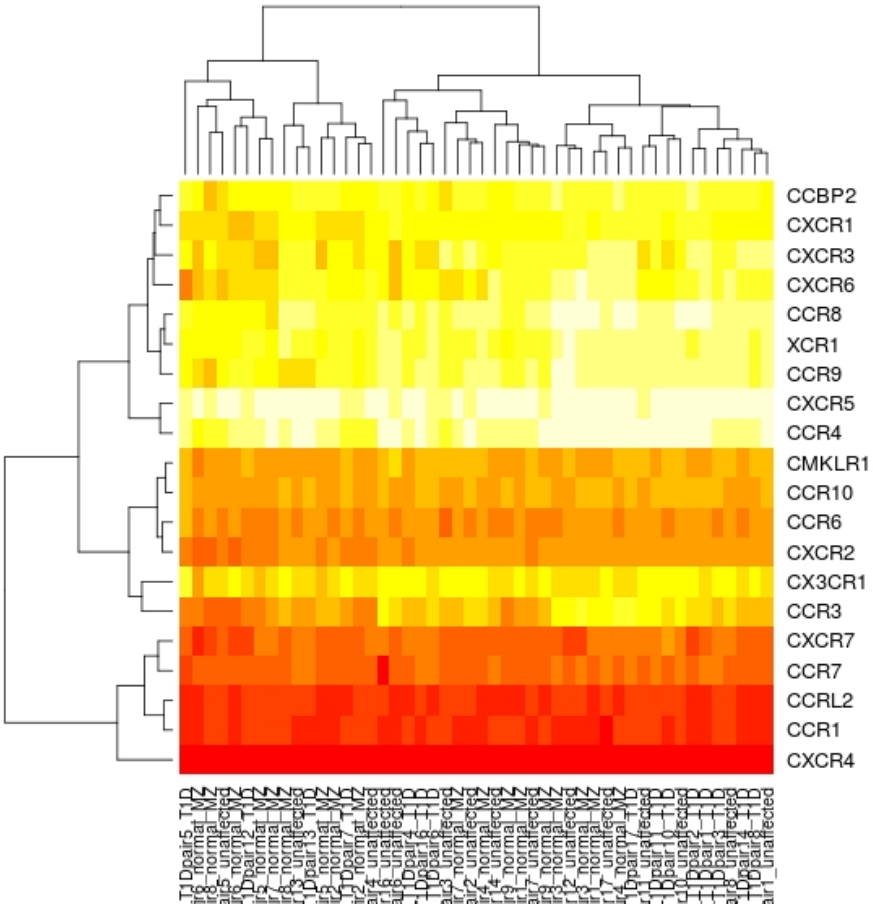


Figure 3.3 Heatmap of the methylation of receptors in Normal and T1D CD4 cells

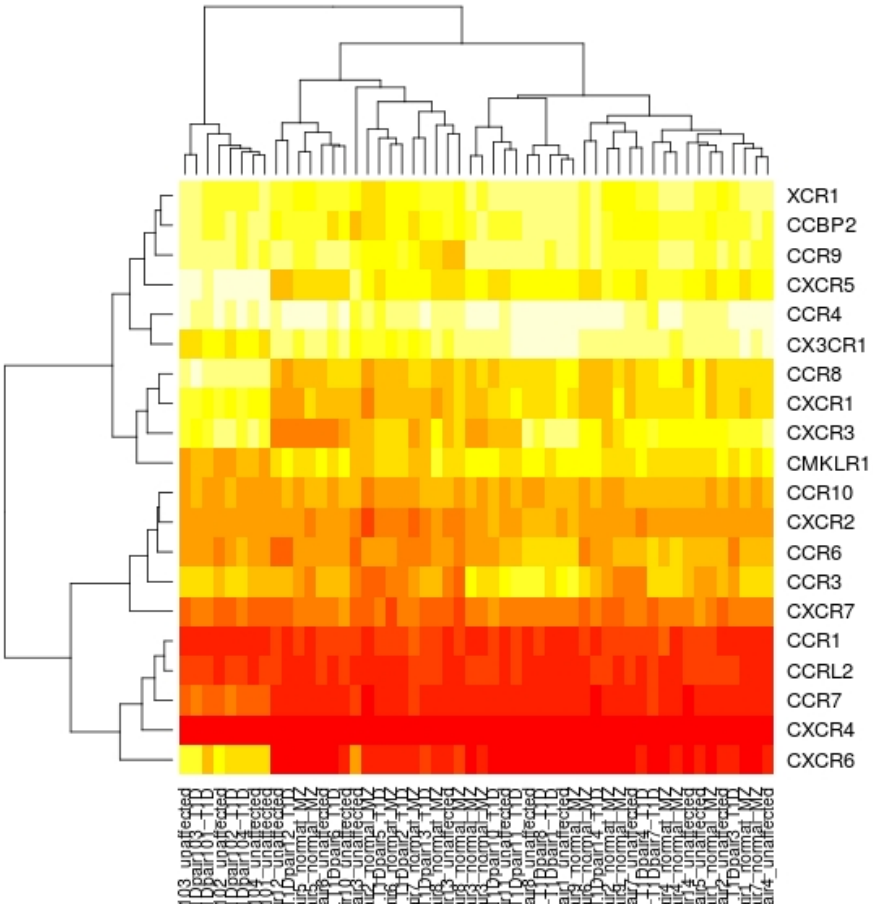


Figure 3.4 Heatmap of the methylation of receptors in Normal and T1D CD14 cells

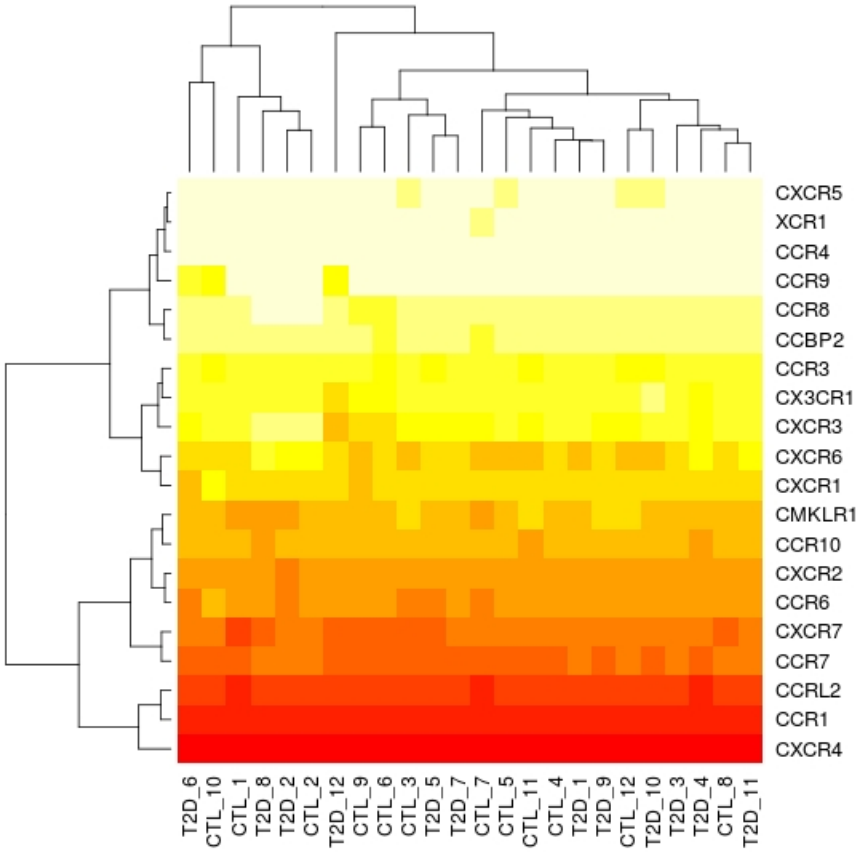


Figure 3.5 Heatmap of the methylation of receptors in Normal and T2D

Statistical analysis didn't show the correlation between methylation and expression. Here, we use a Ornstein Uhlenbeck model to explore the relationships between expression and methylation. We considered 3 adaptive evolution models to describe how phenotypes adapt to each during the phylogeny. We applied these models on the chemokine receptor gene family to explore the potential correlations between expression and methylation. The phylogenetic trees of chemokine receptors are constructed by using MrBayes which is shown in Figure 2.2. The analysis was performed on 3 sample sets:

- T1D samples
- T2D samples
- Healthy samples

The phenotypes considered are the receptor expressions in PBMC and methylation in T1D-cd4, T1D-cd14, T2D-whole blood (depending on availability) measurements.

The main output parameters for the model include ():

a) Matrix A, if A is non-diagonal, there are interactions in the primary optimum.

b) Syy, if Syy is non-diagonal, there are interactions in the stochastic perturbations.

c) mPsi

d) vY0

Output of the best model for healthy samples:

```
A
                Exp       cd4      cd14 wholeBlood
Exp         4.023304  0.00000 0.000000   0.000000
cd4         0.000000 10.03347 0.000000   0.000000
cd14        0.000000  0.00000 3.240734   0.000000
wholeBlood 0.000000  0.00000 0.000000   7.733192

mPsi
               reg.1
Exp         6.3154589
cd4         0.5305160
cd14        0.4647229
wholeBlood 0.5433120

vY0
               [,1]
Exp         6.3154589
cd4         0.5305160
cd14        0.4647229
wholeBlood 0.5433120

Syy
                Exp        cd4       cd14 wholeBlood
Exp         2.158506 -3.2683259 3.7808646 -2.5116383
cd4         0.000000  0.2237032 0.1477688  1.1301598
cd14        0.000000  0.0000000 0.0856523  0.6434418
wholeBlood 0.000000  0.0000000 0.0000000  1.0386175
```

Output of the best model for T1D samples:

```
A
          Exp      cd4      cd14
Exp  10.75385 0.000000 0.000000
cd4   0.00000 7.056942 0.000000
cd14  0.00000 0.000000 2.367516

mPsi
         reg.1
Exp  6.5102267
cd4  0.5367576
cd14 0.4729173

vY0
          [,1]
Exp  6.5102267
cd4  0.5367576
cd14 0.4729173

Syy
           Exp       cd4        cd14
Exp  0.1008103 8.2502759 -5.5318135
cd4  0.0000000 0.1308512  0.9452561
cd14 0.0000000 0.0000000  0.5632444
```

Output of the best model for T2D samples:

```
A
                Exp wholeBlood
Exp        2.591796    0.00000
wholeBlood 0.000000   15.05106

mPsi
               reg.1
Exp        6.4688871
wholeBlood 0.5505259

vY0
               [,1]
Exp        6.4688871
wholeBlood 0.5505259

Syy
                Exp wholeBlood
Exp        3.911532  -3.306315
wholeBlood 0.000000   1.384991
```

All the models for the 3 models have diagonal matrix (A) and all correlations come from the diffusion component (Syy). The estimates of the drift vector (mPsi, optimum value) are similar in all cases. The results suggest that in diabetes the difference between the different conditions was in the diffusion coefficient whilst the drift was similar.

4 Expression and methylation of chemokine receptors

1) DEGs and DMGs

Differentially expressed receptors:
"CCR6" "CCR3" "CXCR5" "CCRL1" "CCR7" "CCR10" "CXCR7" "CMKLR1" "CCR8"
"CCR1" "CXCR3" "CCR5" "CXCR6" "CX3CR1" "CCBP2" "CCR2"

Differentially expressed ligands:
"CXCL13" "CCL19" "CXCL1" "CCL15" "CCL26" "CCL21" "CXCL6" "CCL18" "CXCL2"
"CXCL12" "CXCL11" "CCL20" "CXCL5" "CXCL3" "CCL23" "CCL8" "CCL28" "XCL1" "CCL5"
"CCL24" "CCL11" "CCL16" "CCL4" "CCL13" "XCL2"

Differentially methylated receptors:
"CCR10" "CCR1" "CCR3" "CCR6" "CCR7" "CCR8" "CCR9" "CCRL2" "CMKLR1"
"CX3CR1" "CXCR2" "CXCR3" "CXCR4" "CXCR5" "CXCR7" "XCR1"

Differentially methylated ligands:
"CXCL12" "CCL8" "CXCL1" "CCL7" "CXCL5" "CCL22" "CXCL11" "CCL18" "CCL19"
"CCL20" "CCL27" "CCL5" "CCL2" "CCL23" "CXCL13" "CCL1" "CCL25" "CXCL6" "CCL15"
"CCL13" "CCL11"

Compared with the results in diabetes, more chemokines ligands and receptors are differentially expressed and methylated, and some of them show the same tendencies in the expression and metylation patterns.

2) Correlations between expression and methylation

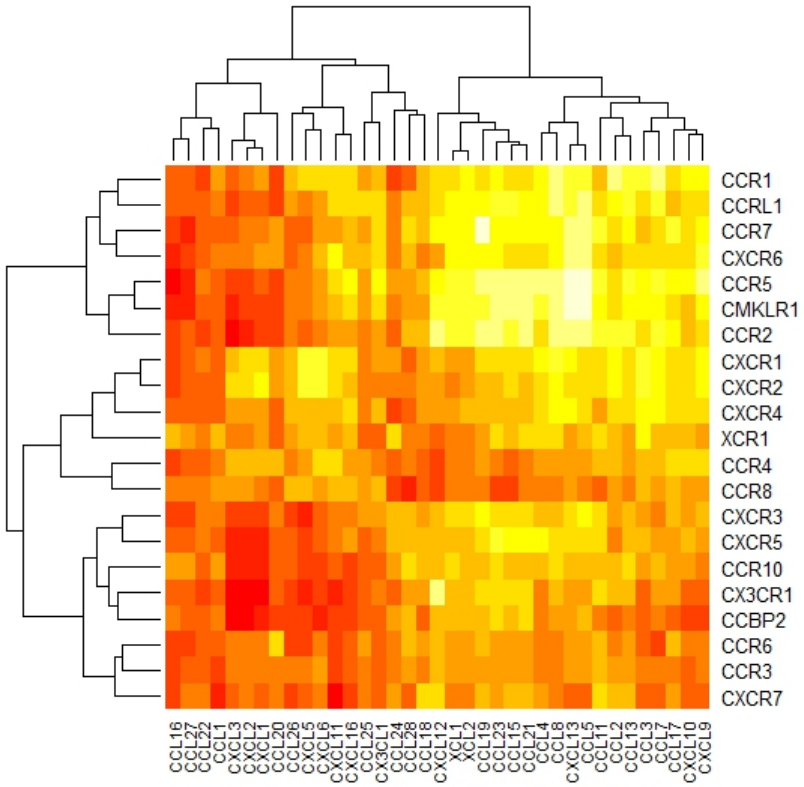


Figure 4.1 Co-expression between ligands and receptors, calculated by Pearson Correlation Coefficient (PCC). The dark color represent low PCC values.

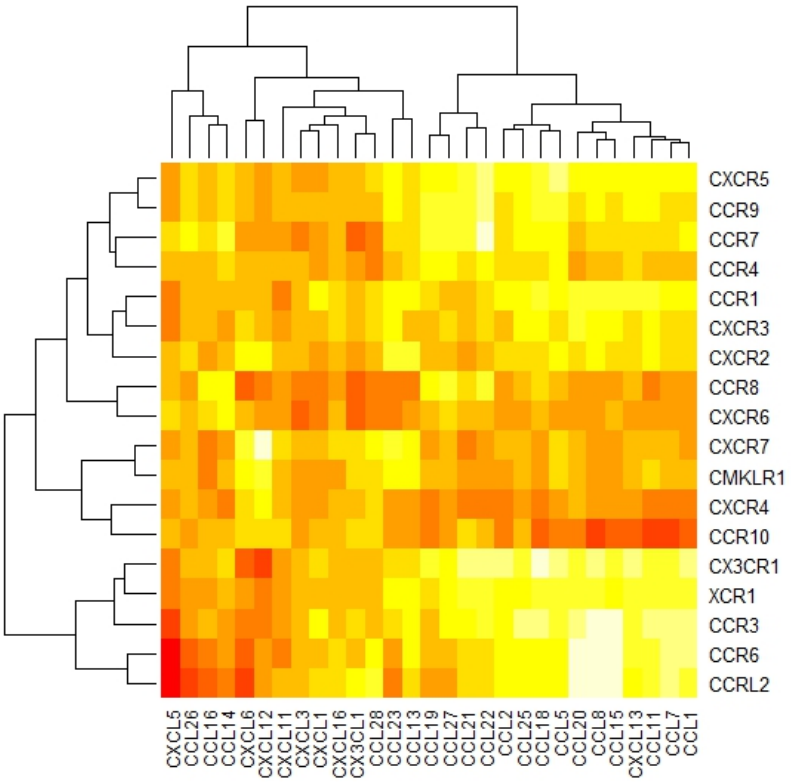


Figure 4.2 Correlations between ligands methylation and receptors methylation, calculated by Pearson Correlation Coefficient (PCC). The dark color represent low PCC values.

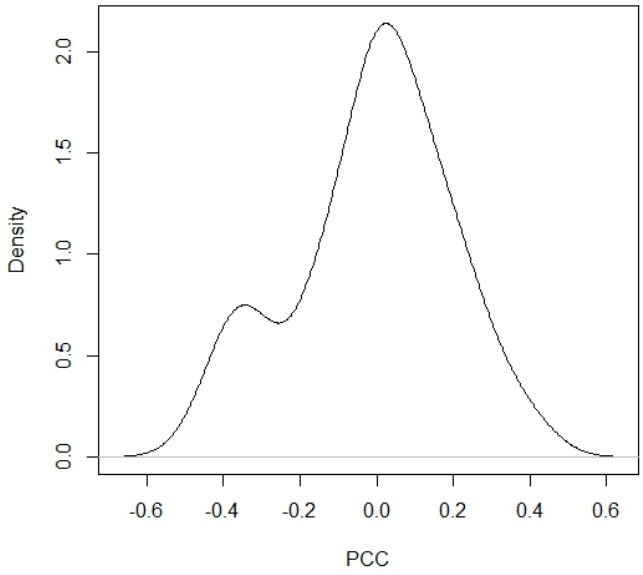


Figure 4.3 Distribution for the PCC between expression and methylation for the receptors/chemokines.

To evaluate the correlations between expression levels and methylation levels, we take the common samples (165) in both expression dataset and methylation dataset of diabetes. The correlations are calculated by PCC, and the distribution is bimodal distribution which shows a small peak in the negative correlations. This is in accordance with the regulation roles of DNA methylation which is mainly repressing the transcription. It suggests the correlations between the expression and methylation of chemokine receptors/ligands in diabetes.

3) multivariate Ornstein Uhlenbeck analysis

The expression and methylation data for diabetes we used were searched and downloaded from GEO database. All of the data are from blood tissue, but there are some differences in the cell lines between the datasets. The expression data is from peripheral blood mononuclear cells. The methylation data for T2D is from the whole blood, while the methylation data for T1D is from purified CD4+ and CD14+ cells from blood. It is reported that epigenetic regulations performs in tissue-specific ways, so the mixed cell lines of the diabetes data may cover the potential correlations between methylation and expression. In the colon diabetes, we just kept the common sample to

calculate the correlations between methylation and correlation, and it shows weak negative correlations. The phylogenetic tree of receptors was also the one constructed by MrBayes. The expression and methylation levels were averaged across all the corresponding samples for diabetes and normal respectively.

The output for best models of Normal samples and inflammatory samples respectively:

Normal samples

```
A
            Exp      Meth
Exp   15.04984 0.000000
Meth   0.00000 3.972027

mPsi
          reg.1
Exp   0.4927272
Meth  0.4834175

vY0
          [,1]
Exp   0.4927272
Meth  0.4834175

Syy
          Exp       Meth
Exp   7.89424 0.0000000
Meth  0.00000 0.8435389
```

inflammatory samples

```
A
            Exp      Meth
Exp   15.04972 0.000000
Meth   0.00000 2.696873

mPsi
          reg.1
Exp   1.1831737
Meth  0.5530236

vY0
          [,1]
Exp   1.1831737
Meth  0.5530236

Syy
            Exp       Meth
Exp   7.418253 0.0000000
Meth  0.000000 0.7694774
```

The diagonal A and Syy indicate that there is not interactions between Exp and Meth, that is, the expression and methylation evolve independently of each other. The statistical analysis suggests that there are some negative correlations between methylation and gene expression. Further final evaluation of the overall data are on going at the time of this report. We will get insights from the collaboration with Anne Ferguson Smith's group (Cambridge).

5 Conclusions

We have developed novel algorithms and pipelines that allow to integrate several omics data (CVN, methylation, gene expression). The large amount of data we are producing will be interpreted using artificial intelligence programs. Two papers have been completed (in submission) and we hope to have more.

6 Bibliography

Bae JS, Cheong HS, Kim JH, Park BL, Kim JH, et al. (2011) The genetic effect of copy number variations on the risk of type 2 diabetes in a korean population. PloS one 6: e19091

McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nature genetics 39: S37-S42.

Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5: R80.

Therneau T, Ballman K (2008) What does plier really do? Cancer informatics 6: 423.

Bartoszek K., J. Pienaar, P. Mostad, S. Andersson and T. F. Hansen. (2012) A phylogenetic comparative method for studying multivariate adaptation. J. Theor. Biol., 314:204–215.

Butler M.A and A. A. King A.A. (2004). Phylogenetic comparative analysis: a modelling approach for adaptive evolution. Am. Nat., 164(6):683–695.