Genome-wide association studies for CMT1A patients-specific methylation patterns using python open source code

Seon Hyeok Hwang[p1], Jun Yeop Yim[2], Da Eun Nam[1], Hee Ji Choi[1], Jin Hee Park[1], Young Jin Park[1], Byung Kwon Pi[1], Byung-Ok Choic[3], Ki Wha Chung[c1]

[1]Department of Biological Sciences, Kongju National University, 56 Gongjudaehak-ro, Gongju 32588;
[2]Department of Applied Mathematics, Kongju National University, 56 Gongjudaehak-ro, Gongju 32588;
[3]Department of Neurology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351

KONGJU NATIONAL UNIV.
APPLIED GENETICS LABORATORY

Abstract

Charcot-Marie-Tooth disease type 1A (CMT1A) is caused by 17p12 region duplication, but clinical heterogeneity ranges from mild to severe. This study was performed to determine whether epigenetic factors affected to severity. The subjects were investigated in 11 unaffected individuals (42.2 ± 3.4 years old) and 22 CMT1A patients (46.4 ± 3.5 years old) in male. Using the SureSelectXT Methyl-Seq Library Kit, the methylation levels of a total of 6,279,954 CpG sites were measured. In order to apply an appropriate statistical method according to the data characteristics of each CpG site, we coded to statistical program using pandas, a python-based open source library. CpG sites selected less than 10 significant p-values (P < 1.00E-05). In addition, a significant hypomethylation pattern of 1.1 kb was found in the 17p12 region of the CMT1A patient group (P < 0.05). This 1.1 kb intergenic region is between the genes cytochrome c-oxidase 10 (*COX10*) and CMT1A duplicated region transcript 15 (*CDRT15*) in the CMT1A-REP. Therefore, Duplication of the 17p12 region resulted in this methylation level. This results will help develop biomarkers or personalized medicine. In addition, it is expected that bioinformatics will be universalized through the application of personal-scale statistical programming

Introduction & Purpose

- Methylation or demethylation are universally affected to repression and promotion in expression of genes.
- It has been reported that inhereditary neuropathy patients, such as SMA show severity differences according to methylation patterns.
- Methylation analysis between CMT1A patients group and unaffected individuals group, was performed to find epigenetic differences.

Subjects & Methods

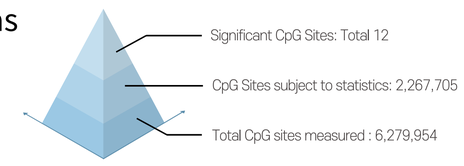
- The epigenome wide association study (EWAS) was performed in 22 CMT1A patients and 11 unaffected individuals in male.

22 CMT1A patients group (46.4 ± 3.5 years old) **11 unaffected individuals (42.2 ± 3.4 years old)**

	FC272-7	FC339-1	FC502-1	FC296-2	FC492-1	FC207-1	FC403-1	FC434-1	FC709-1	FC617-3	FC380-1	FC181-1	FC546-1	FC738-2	FC543-1	FC719-1	FC604-3	FC296-1	FC309-1	FC607-1	FC471-8	FC1178	FC113-4	FC189-2	FC256-3	FC316-2	FC317-03	FC394-3	FC433-3	FC476-3	FC640-4	FC701-10	FC732-2
sex	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M
age	42	44	46	55	44	52	47	47	45	47	42	48	49	39	45	46	47	50	49	43	47	47	42	45	38	37	45	41	42	45	47	38	44
onset age	8	8	.	11	37	19	25	15	.	12	5	47	40	15	10	13	5	20	40	39	.	18	.	.	.	.	.	.	.	.	.	.	.
exam age	35	37	46	47	40	43	40	42	45	45	36	48	41	39	42	45	45	43	42	40	47	47	.	.	.	.	.	.	.	.	.	.	.
Disease duration	27	29	.	36	3	24	15	27	.	33	31	1	1	24	32	32	40	23	2	1	.	29	.	.	.	.	.	.	.	.	.	.	.
CMTNS	23	20	18	17	17	16	15	15	15	14	14	14	12	12	11	11	10	9	9	9	6	6	.	.	.	.	.	.	.	.	.	.	.

Table 1. Clinical information of CMT1A patients group and uunaffected individuals group.

- Methylation measure was used by SureSelectXT Methyl-Seq Library Kit based on the reference sequence hg19, and the lower read (500) and higher read cutoffs (10) were excluded.
- Statistical analysis program was coded with the python based open source library pandas.




```
for methylation levels in all positions in chromosome do

    let A be a list of float of methylation level of patient
    let B be a list of float of methylation level of non-patient
    let p-value be a float of Levene-test of (A,B)

    if p-value > 0.05 then
        set Student's t-test ← equal variance
        let result of t-test be a float of p-value of Student's t-test of (A,B)
    else
        set Student's t-test ← non-equal variance
        let result of t-test be a float of p-value of Student's t-test of (A,B)
    let result be a list of position, A, B, p-value, result of t-test
return all result
```

Figure 1. Pseudo code for statistic algorithm.
The algorithm was coded with Python-based library Pandas.

python™ Python — Using Python, for programming language

pandas Pandas — The source code referred to Pandas that Python-based programming software library

SCAN ME!

If you want to see more detailed codes, please refer to this QR code.

Results

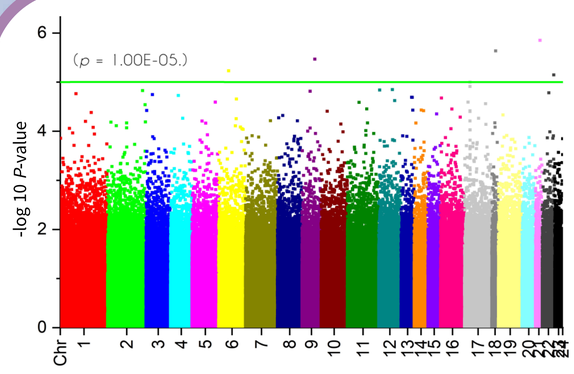


Figure 2. Manhattan plot of Student's T-test p-values.

Test statistic	Region	Gene	Chr	Position	*P*-value
Student's T-test	intergenic	C21orf33, ICOSLG	21	45580961	1.41E-06
	intergenic	MC4R, CDH20	18	59001216	2.30E-06
	exonic	PRRC2B	9	134351466	3.39E-06
	ncRNA_exonic	LOC285847	6	35697128	5.88E-06
	intronic	MOV10L1	22	50585136	7.13E-06
Welch's T-test	intergenic	CRADD, PLXNC1	12	94376893	2.19E-06
	intergenic	WDR1, ZNF518B	4	10294645	3.46E-06
	intronic	SPOPL	2	139307145	4.74E-06
	exonic	PACSIN1	6	34495259	5.56E-06
	intronic	KDM4B	19	5034228	7.82E-06
Mann-Whitney U test	intronic	CDHR2	5	175970622	2.21E-06
	intronic	CDHR2	5	175970657	5.49E-06

Table 2. Significant p-value for each test (P < 1.00E-05).

CpG sites information				CMT1A Patients group		Controls group		p-value
Region	Gene	Chr	Position	Average	Standard Deviation	Average	Standard Deviation	
intergenic	COX10, CDRT15	17	14114556	0.964	0.015	0.974	0.011	0.030516
intergenic	COX10, CDRT15	17	14115285	0.252	0.111	0.369	0.166	0.004367
intergenic	COX10, CDRT15	17	14115299	0.292	0.138	0.435	0.194	0.005262
intergenic	COX10, CDRT15	17	14115314	0.325	0.123	0.475	0.159	0.000556
intergenic	COX10, CDRT15	17	14115358	0.379	0.118	0.542	0.168	0.000228
intergenic	COX10, CDRT15	17	14115370	0.352	0.109	0.524	0.171	9.38E-05
intergenic	COX10, CDRT15	17	14115388	0.425	0.095	0.587	0.162	3.83E-05
intergenic	COX10, CDRT15	17	14115393	0.422	0.093	0.595	0.166	1.23E-05
intergenic	COX10, CDRT15	17	14115396	0.393	0.093	0.566	0.160	1.00E-05
intergenic	COX10, CDRT15	17	14115418	0.485	0.095	0.624	0.135	6.16E-05
intergenic	COX10, CDRT15	17	14115449	0.549	0.090	0.669	0.162	0.000204
intergenic	COX10, CDRT15	17	14115470	0.308	0.075	0.434	0.129	3.96E-05
intergenic	COX10, CDRT15	17	14115539	0.856	0.043	0.890	0.033	0.005866
intergenic	COX10, CDRT15	17	14115564	0.683	0.073	0.773	0.105	0.000339
intergenic	COX10, CDRT15	17	14115656	0.831	0.055	0.887	0.064	0.003109
intergenic	COX10, CDRT15	17	14115676	0.954	0.022	0.970	0.018	0.02123

Table 3. Significant hypomethylation region on 17p12 (P < 0.05).

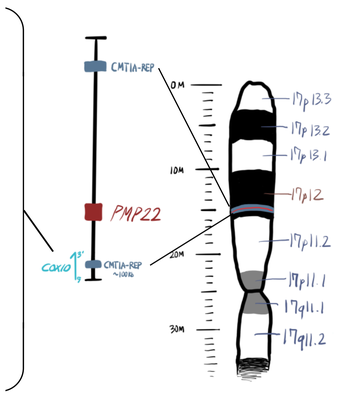


Figure 3. The genetic locus of COX10

GO biological process complete	REFLIST (20589)	UPLOAD (699)	expected	fold Enrichment	P-value	FDR
system development	3838	208	130.3	1.6	3.26E-12	5.11E-08
cell differentiation	3519	194	119.47	1.62	5.47E-12	4.29E-08
anatomical structure morphogenesis	2237	139	75.95	1.83	7.16E-12	3.74E-08
multicellular organism development	4228	222	143.54	1.55	9.51E-12	3.73E-08
cellular developmental process	3542	194	120.25	1.61	1.20E-11	3.75E-08
anatomical structure development	5144	254	174.64	1.45	6.38E-11	1.67E-07
developmental process	5677	274	192.74	1.42	7.67E-11	1.72E-07
nervous system development	2191	133	74.38	1.79	1.33E-10	2.60E-07
neurogenesis	1290	88	43.8	2.01	1.68E-09	2.92E-06
animal organ development	3254	171	110.47	1.55	6.88E-09	1.08E-05

Table 4. Gene Ontology Resource search results of genes with P < 0.001 sites in Student's t test.

Conclusions

- A total of 12 sites showed significant differences between patients and unaffected individuals (P < 1.00E-05), and methylation patterns of the 1.1 kb region in 17p12 were hypo in the CMT1A group (P < 0.05).
- In the biological process analysis, a total of 689 genes with sites of P < 0.001 showed high association in neurons or neurons.
- This study is expected to be helpful for biomarker development or personalized medicine research.