의학석사 학위논문

Somatic mutations during reprogramming to pluripotency revealed by massively parallel sequencing of iPS cell whole genomes

전장유전체서열 분석을 통한 만능유도 줄기세포의 체세포 변이 연구

2013 년 02 월 05 일

서울대학교 대학원

의학과 석사과정

김 유 린

논문 제목: 전장유전체서열 분석을 통한 만능유도 줄기세포의 체세포 변이 연구

학위구분: **석사 ■ · 박사 □**
학　　과: 의과학과
학　　번: 2011-23019
연 락 처:
저 작 자: 김유린　　　　　(인)

제 출 일: 2013 년　02 월

서울대학교총장 귀하

ABSTRACT

Background: Despite induced pluripotent stem cell (iPSC) reprogramming success, the whole genome sequence of iPSC has rarely been explained at high throughput. As iPSC can potentially replace embryonic stem cells for clinical applications as a therapeutic agent it is important to understand genomic changes occurring during the reprogramming process. As described by Nagy A. *et al.* iPS cell fate is uncertain during a timeframe from Day 5 to Day 15, and reprogramming to pluripotency is unavoidable after Day 16. In order to assess mechanisms necessary to the reprogramming and arising genomic aberrations, we studied variants arising before, within and after the timeframe of uncertain cell fate in iPSC generated from mouse fibroblasts.

Method: We generated a whole genome sequence with the next generation sequencer Illumina HighSeq 2000 at a high coverage of 35.77X in average of 5 samples; primary iPSC, three intermediate stages of reprogramming at Day 0, Day 11 and Day 18, and secondary iPSC. We then performed single nucleotide polymorphism and short insertion-deletion genotyping with our GMI caller and compared our method with two different algorithms, Samtools and MAQ. Subsequent gene ontology was performed with Ontologizer 2.0.

Results: We identified more than 5 million single nucleotide polymorphisms (SNPs) per sample and our results showed that the number of non-synonymous SNPs, unique SNPs per sample compared to 1iPSC and the number of indels reached a maximum at D11 and then decreased until 2iPSC. Thus as reported by Ji J. *et al.* we found that cells reprogramming to pluripotency have an increased number of somatic coding mutation from D11 to 2iPSC. Gene Ontology of somatic non-synonymous SNPs occurring at the end of the reprogramming revealed that biological process such as gene expression ($p=9.05 \times 10^{-3}$), cell differentiation ($p=9.38 \times 10^{-3}$) and cell proliferation ($p=7.79 \times 10^{-3}$) are affected by coding mutations. These results suggest that genes involved in these biological processes might have a role in the reprogramming to pluripotency.

Conclusion: The timeframe of uncertain cell fate is a critical time limit for the reprogramming process. Like previously reported by Ji J. *et al.* our results showed an elevated coding mutation rate during the reprogramming. Finally, gene ontology showed that coding mutations are affecting important biological process at the end of the reprogramming such as cell differentiation, the immune system and gene expression. Therefore, we encourage further characterization of these variants to understand mechanisms involved during the timeframe of uncertain cell fate.

Keywords: Induced pluripotent stem cell, stem cell, reprogramming, *Yamanaka* four factors, whole genome sequencing, single nucleotide polymorphism, insertion-deletion.

Student number: 2011-23019

CONTENT

ABSTRACT i
CONTENT iv
LIST OF TABLES AND FIGURES v
LIST OF ABBREVIATIONS 1
INTRODUCTION 2
1. Reprogramming of mouse embryonic fibroblast to induced pluripotent stem cell 2
1.1 Reprogramming of mouse embryonic fibroblast with the four Yamanaka factors 3
1.2 Timeframe of uncertain cell fate during the reprogramming to pluripotency 6
MATERIALS AND METHODS 7
1. Reprogramming to iPSC with the four *Yamanaka* factors 7
2. Massive parallel sequencing of iPS cell Whole Genomes 8
3. SNP and insertion-deletion genotyping of samples undergoing reprogramming to pluripotency and iPSCs 9
4. Validation of single nucleotide polymorphism 11
RESULTS 15
1. Whole Genome sequencing of fibroblasts reprogramming to pluripotency 15
1.1. Validation and comparison with two different algorithms for single nucleotide polymorphisms discovered in fibroblasts reprogramming to pluripotency 17
2. Strain differences and homozygosity rate of the five samples 21
2.1.Number of SNPs within gene boundaries and nsSNPs per sample 23
2.2.Number of unique SNPs per sample after comparison with 1iPSC 25
2.3. Somatic SNPs arising and maintained during the reprogramming to pluripotency 27
2.3.1.Gene Ontology of somatic nsSNPs arising and maintained during the reprogramming 30
2.4. SNPs found within miRNA of 1iPSC and 2iPSC 36
3. Indels found within the five samples 38
DISCUSSION 41
CONCLUSION 46
ACKNOWLEDGEMENT 47
국문 초록 49
REFERENCES 51

LIST OF TABLES AND FIGURES

Figure 1. Schematic representation of the reprogramation of a mouse embryonic fibroblast with the four *Yamanaka* factors (*c-Myc*, *Klf4*, *Oct4*, *Sox2*)..4
Figure 2. Schematic representation of the timeframe of uncertain cell fate as described by Nagy A. *et al.*[6]......................................5
Figure 3. Screenshot of SNPs selected randomly and validated by PCR and capillary sequencing..19
Figure 4. Ratio of homozygosity and heterozygosity throughout the reprogramming from 1iPSC to 2iPSC..22
Figure 5. Number of unique SNPs per samples after comparison with 1iPSC...26
Figure 6. Number of somatic nsSNPs arising from D0 to 2iPSC.....29
Figure 7. Representation of GO terms found by gene ontology for somatic nsSNPs arising between D18 and 2iPSC......................32
Figure 8. Representation of the number of indels causing frameshifts in the genome. 40

Table 1. Primer sequences of validation by capillary sequencing of 1iPSC, D0, D11, D18 and 2iPSC of homozygous SNPs. 13
Table 2. Primer sequences of SNPs randomly selected among the five samples...14
Table 3. Sequencing summary of 1iPSC, D0, D11, D18 and 2iPSC (GSNAP and BWA). 16
Table 4. Summary of SNP genotyping with GMI caller, Samtools and MAQ. 20
Table 5. Summary of SNPs found within gene boundaries and nsSNPs at each step of the reprogramming to pluripotency....24
Table 6. Identification of somatic SNPs maintained during the reprogramming from D0 to 2iPSC...28
Table 7. Gene ontology summary of nsSNPs appearing at D11. ..33
Table 8. Gene ontology summary of nsSNPs appearing at D18. ..34
Table 9. Gene ontology summary of nsSNPs appearing at 2iPSC.35
Table 10. Summary of SNPs found in gene coding for miRNA....37
Table 11. Summary of the number of indels found within the five iPSC...39

LIST OF ABBREVIATIONS

1iPSC	Primary iPSC
2iPSC	Secondary iPSC
bp	Base Pair
D0	Day 0
D11	Day 11
D18	Day 18
GO	Gene Ontology
iPSC	induced Pluripotent Stem Cell
MEF	Mouse Embryonic Fibroblast
nsSNP	non synonymous SNP
sSNP	synonymous SNP
SNP	Single Nucleotide Polymorphism
WGS	Whole Genome Sequencing

INTRODUCTION

1. Reprogramming of mouse embryonic fibroblast to induced pluripotent stem cell

Induced pluripotent stem cells (iPSCs) are pluripotent stem cells artificially derived from a non-pluripotent adult somatic cell by an inducing method with specific genes. The iPSCs have the same anatomical characteristics as stem cells and the two are said to be virtually indistinguishable from each other[1]. Stem cells represent an auspicious therapy in regenerative medicine; however, bioethical issues limit the use of fertilized eggs to create stem cells and patients receiving a stem cell treatment must be given immunosuppressant drug to prevent rejection, leading to a low success of this therapy. An alternative to stem cells is iPSC. The latter can be induced from the patient's fibroblast, rending the use of immunosuppressant drug unnecessary, thus no fertilized egg is necessary to obtain iPSC which discards any bioethical issue[2].

The iPSCs are method-sensitive and a successful reprogramming depends highly on the template used for reprogramation. Various techniques can lead to a successful reprogramming. Among all methods available to reprogram a cell into a pluripotent stem cell, direct reprogramming is the most efficient technique, such as direct reprogramming by miRNA[3], direct by the use of unique gene[4], or direct reprogramming by the use of the four *Yamanaka* factors[5].

1.1 Reprogramming of mouse embryonic fibroblast with the four Yamanaka factors

Reprogramming with the four *Yamanaka* factors allows an embryonic fibroblast to reach a pluripotent state and does not require the use of fertilized eggs. It is processed through the transfection of a vector containing the four *Yamanaka* factors (*c-Myc*, *Klf4*, *Oct4*, *Sox2*) in mouse embryonic fibroblasts (MEFs). In our study, MEFs were transfected with the PB-TET vector (Figure 1) under the presence of doxycycline. As this vector is doxycycline dependant, transcription of the four factors only starts when placed in a feeder environment with doxycycline.

As shown in Figure 2, a first line of reprogramming has been done from MEF to a first generation of iPSC (1iPSC). The 1iPSCs obtained were then aggregated with a diploid or a tetraploid chimera, giving a second generation of MEF. The secondary MEFs were transfected in a second step, placed under doxycycline environment and underwent through a second reprogramming yielding our second generation of iPSC (2iPSC).

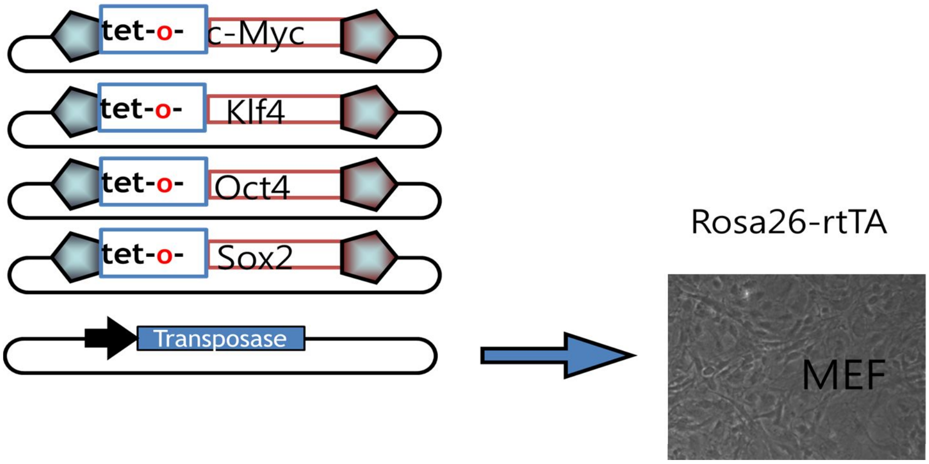


Figure 1. Schematic representation of the reprogramation of a mouse embryonic fibroblast with the four *Yamanaka* factors (*c-Myc*, *Klf4*, *Oct4, Sox2*).

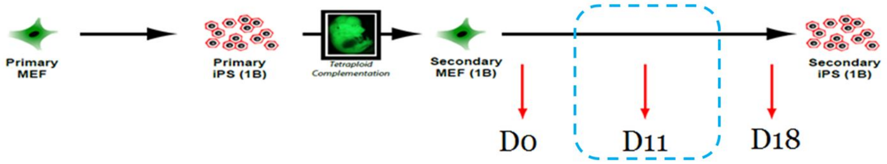


Figure 2. Schematic representation of the timeframe of uncertain cell fate as described by Nagy A. *et al.*[6] The time frame of uncertain cell fate starts from Day 8 to Day 15 and is represented here by a dotted blue line.

1.2 Timeframe of uncertain cell fate during the reprogramming to pluripotency

The reprogramming to pluripotency is divided into three phases, as described by Nagy A. *et al.*[6] The first phase is a timeframe of about 8 days, during this period of time, removing reprogramming cells from their feeder environment causes a break in the reprogramming and cells return to their somatic state. After Day 16, even if cells are removed from their feeder environment they are still fully reprogramming toward pluripotency. However during a time-window from Day 8 to 15, cells fate is undetermined and could as well return to a somatic state or move toward pluripotency. We call this time-window the *Timeframe of uncertain cell fate*, referenced as *Area 51* by Nagy A. *et al.*[6]. Mechanisms responsible for cell fate and reprogramming occurring during the timeframe of uncertain cell fate have yet been established clearly. The therapeutical power of iPSC is undoubtedly applicable to patients in regenerative medicine if science is able to offer stable cells and able to control their reprogramation and differentiation. In order to understand mechanisms underlying the reprogramation process to pluripotency, we performed whole genome sequencing (WGS) at high throughput of the first and second generation of iPSC and 3 samples before, during and after the timeframe of uncertain cell fate.

MATERIALS AND METHODS

An extensive review of the materials and methods cannot be done without a clear explanation of the provenance and reprogramation methods used to obtain the two iPSC as the method could influence considerably the output of these results. We handled in our facility WGS at high throughput of the five samples and variant genotyping; therefore part 1 of the present materials and methods is only to provide the reader with all information necessary to understand the reprogramation method.

1. Reprogramming to iPSC with the four *Yamanaka* factors

The mouse strain used for this experiment is a hybrid of two 129 sub-strains 129X1/SvJ and 129S1/Sv-Oca2+ Tyr+ KitlSl-J and the strain C57BL/6J. Fibroblasts were induced to 1iPSC by Piggyback transposition with the four *Yamanaka* factors (*c-Myc*, *Klf4*, *Oct4*, *Sox2*) according to the *Yamanaka* protocols[2], and aggregated to a tetraploid chimera. Then embryos were derived to a secondary MEF and underwent through a second reprogramming to the 2iPSC. Between the secondary MEF and 2iPSC, three intermediate stage were removed from the feeder environment at Day 0 (D0), Day 11 (D11) and Day 18 (D18) and constituted our three intermediate samples between 1iPSC and 2iPSC.

2. Massive parallel sequencing of iPS cell Whole Genomes

Sequencing library were generated from genomic DNA of five samples, 1iPSC, D0, D11, D18 and 2iPSC, furthermore paired-end WGS with the next generation sequencer Illumina HiSeq 2000 was performed according to the manufacturer's protocol and our previous reports[7,8]. The DNA insert size was 275bp and the read length was 100bp. Reads having a Phred scores of 20 and higher were selected and aligned to the mouse reference genome mm9 NCBI build 37 with the aligner GSNAP[9] and 5% mismatches were allowed. Percent coverage of the genome was calculated for ≥1x.

The second line of alignment has been performed with two different algorithms, BWA and MAQ. We aligned our 5 samples (1iPSC, D0, D11, D18 and 2iPSC) with BWA [10] v0.5.9-r16 (http://bio-bwa.sourceforge.net/). PCR duplicates were removed with PICARD v1.55 (http://picard.sourceforge.net/) and sequencing data analyses were performed with GATK [11] v1.4.11 (http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit). We also performed the third alignment and SNP call with MAQ v0.6.8 (http://maq.sourceforge.net) for two samples, D0 and 2iPSC.

3. SNP and insertion-deletion genotyping of samples undergoing reprogramming to pluripotency and iPSCs

From sequences aligned with GSNAP, SNPs were called with our GMI SNP caller. SNP was called if 4 or more unique reads aligned to the said SNP, 20% or higher of total aligned reads aligned to the SNP and an average quality score of 20 or higher was present.

The percent of aligned reads to the SNP is also a criterion to determine heterozygosity or homozygosity of the SNP. A SNP is said heterozygote if the number of reads aligned to it accounts from 20% up to 89%; and homozygote when 90% or more reads aligned to the SNP.

From sequences of 1iPSC, D0, D11, D18 and 2iPSC aligned with BWA, variants were called with Samtools [12] v0.1.18 (http://samtools.sourceforge.net/). From sequences of D0 and 2iPSC aligned with MAQ, SNPs were called with MAQ v0.6.8 (http://maq.sourceforge.net). We extracted the common list of SNPs called between Samtools and GMI caller in order to proceed further with the analysis.

For the total SNPs of each sample, we annotated them with the NCBI RefSeq gene set. SNPs falling within gene boundaries were then classified as being non-synonymous SNPs (nsSNPs) or synonymous SNPs (sSNPs).

The number of unique SNP was calculated by comparison with the number of SNPs with 1iPSC for each sample. Therefore, the number of unique SNPs represents the number of SNPs that were found in the sample but not in 1iPSC.

As our study focus on the outcome of the timeframe of uncertain cell fate we studied SNPs arising and maintained during the reprogramming. A SNP is said *somatic* if it arises during the reprogramming, *i.e.* it hasn't been identified in 1iPSC but is present in 2iPSC. *A contrario* a SNP is called *germline* if it is inherited from the parent cell line *i.e.* present all through the reprogramming from 1iPSC to 2iPSC. We have identified somatic SNPs occurring during the reprogramming at D0, D11, D18 and 2iPSC. By comparison with coverage data, a SNP is considered somatic if 10 or more reads aligned to the reference, for example at D0, and no read aligned to the SNP, plus, 4 reads or more aligned to the SNP form, for example at D11. Every somatic SNP falling within gene boundaries were classified as synonymous or non-synonymous.

Finally, gene ontology was performed on somatic every nsSNP with a multifunctional tool for GO term enrichment analysis, Ontologizer v2.0[13].

From a common list of SNPs between 1iPSC and 2iPSC SNPs within boundaries of genes coding for miRNA were extracted. We counted the number of SNPs within each gene coding for a miRNA. Also, SNP is said *novel* if it hasn't been classified in dbSNP128[14].

Indels were called with our GMI indels caller. By comparison with the coverage data, an indel is called when when 4 or more unique reads aligned to this variation, 20% of higher reads aligned to the indel and an average quality of 20 or higher is present. An indel is said heterozygote if the number of reads aligned to the indel accounts for 20% up to 59%; homozygote when 60% or more reads aligned to the indel.

4. Validation of single nucleotide polymorphism

Variants found were validated by PCR and Capillary Sequencing. We validated our GMI SNP call by randomly selecting 10 homozygous germline SNPs. In addition, we randomly selected 13 SNPs in various samples and determined if homozygosity and heterozygosity has been called correctly and if the allele were concordant. We then calculated the total concordance rate of our GMI SNP caller.

Table 1. Primer sequences of validation by capillary sequencing of 1iPSC, D0, D11, D18 and 2iPSC of homozygous SNPs.

Chr.[a]	Pos.[b]	Left primer	Right primer
2	111620649	TTTTTATTTCAGGTGCCGATG	CCAGGAAGTCAACACCAACC
2	156574134	CAAAAGCGATCGATGTGATG	GGAGACAGAGAAACGGATGC
4	117687333	GGCCAGTCAAGCACATTCTA	AGCCGTAGAGCCTTTGTGGT
4	140773318	TGACAAGCAGATCAACTGTGAG	TGTGGCCACCACTGTCTCTA
7	110701107	CAGACTGTGCTCTGCCTCAG	CGACACAGAGTTAGATCCCTCA
8	23671603	TGGTTTTGATGGACCAGAGC	GGATTAAGACACTGACCTCACG
8	44275679	CAGGGAGAGAAGAGGGTGAG	GGAACACTAGATAAAGGAATTACTCG
8	71597682	TGTTTCTTCACAGCAATAGAGCA	AAGCATTGATGCCCAGGTTA
9	105354198	GGCAAAAAGGAAACATGGAA	CCTCCAAGACTCCAGGTTATG
11	70995526	CCACAATTCTGTCCATTTCCA	GACCAGAATCCTGAGCTATGTCT

[a] Chr. = Chromosome

[b] Pos. = Position

Table 2. Primer sequences of SNPs randomly selected among the five samples.

Sample	Chr.[a]	Pos.[b]	Left primer	Right primer
1iPSC	2	146853583	CTTGGCAAGTTGTGGTTCTTC	CATGTGTAGCAAGACCAAGCA
	3	155752942	CATTTTTGCAAAGAAGGAAAAAG	AAGCAAAGGTGGATCCTGAA
	6	135414617	TGCATAAGTATGTGCCAGGTG	GCTTCCCAGGAATGTCTGAA
	19	24806112	AGCAAATGTGTTCAGACATGG	AGTTGGAGGTTGTGCTAGGG
D0	2	70810683	GCAGGTCTGTACCCTCTTCCT	TCACCCGAGTCCAGTGAGA
	5	7578313	CCCATTCCACATTGTCTTCC	TGGAAGCAGAAAGAAAAGAACA
D11	1	135017429	AGGGCCAAGAGCGAAAATAA	TACAGGGATTTGGAGCCTGA
	4	130614543	ACCAGTCCCATGTTGTAGCC	ACCATACCGTGGCATTCATT
D18	2	148822433	TCAACCAGGTGAGGTCTTCTC	CACAGCTGAGCAATTGTCTTTT
	8	4953252	TAAATTGCCAACAGCCAAAG	AATCCATTTAAAAATCACCACAG
	10	26337612	AACTGCCAAGTTTACAGCTGAG	TCCCACACCCTGAGAGGTTA
2iPSC	1	136589618	GTAGTACTGGGCTGGGCTGA	TCCCTTCTCAAAGGCTTTCC
	1	145853863	AGGGGCCCTGGATTCTATCT	TTGGCTTTAAATTACTAGAGCATGA

[a] Chr. = Chromosome

[b] Pos. = Position

RESULTS

1. Whole Genome sequencing of fibroblasts reprogramming to pluripotency

According to our alignment with GSNAP, the percent of whole genome of fibroblast reprogramming to pluripotency, which is covered at ≥1x, ranges from 97.58% to 98.39%. The maximum coverage depth was 42.80x at D11 and the average coverage depth of the five samples was 35.77x (Table 3). Whereas, alignment with BWA covered the whole genome at ≥1x from 91.75% to 92.81%, and reached a maximum coverage depth of 36.74x at D11. The average coverage depth was 27.01x. We can see that both alignments follow the same pattern with lower coverage depths for 1iPSC and 2iPSC and a maximum coverage depth at D11.

Table 3 . Sequencing summary of 1iPSC, D0, D11, D18 and 2iPSC (GSNAP and BWA).

		GSNAP			BWA		
Sample	Total number of reads	Aligned reads	Coverage depth (x)	Percent of the genome covered (%)	Aligned reads	Coverage depth (x)	Percent of the genome covered (%)
1iPSC	943,825,850	793,686,800	31.02	97.63	488,036,626	19.07	91.76
D0	1,541,722,498	1,095,095,527	42.80	98.33	843,050,303	32.95	91.75
D11	1,533,460,484	1,032,761,914	40.37	98.39	939,918,192	36.74	92.76
D18	1,217,252,574	873,128,378	34.13	97.89	724,541,780	28.32	92.81
2iPSC	950,272,858	781,684,533	30.55	97.58	460,095,965	17.98	92.27

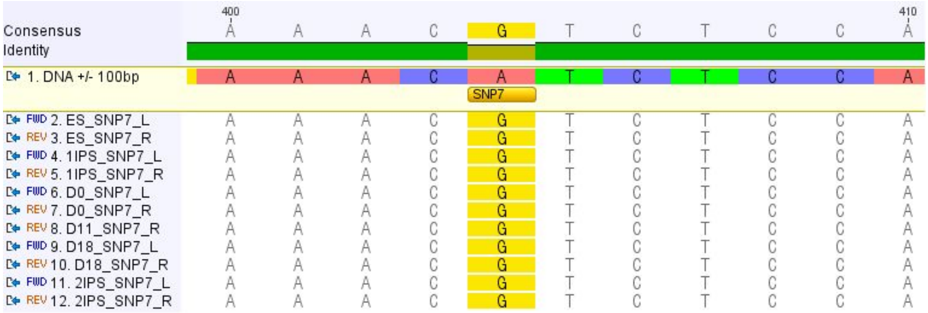
1.1. Validation and comparison with two different algorithms for single nucleotide polymorphisms discovered in fibroblasts reprogramming to pluripotency

Among the SNPs randomly selected, 10 SNPs were from a list of homozygous SNPs found in all 5 samples and were validated by PCR and capillary sequencing (Figure 3). Out of 10 SNPs selected 9 of them were concordant which let us conclude that for homozygote SNP genotyping the concordance of our GMI SNP call is 90% for homozygous SNPs.

In the second part of the validation, we randomly selected 13 SNPs and performed capillary sequencing. Out of 13 SNPs selected 11 of them were concordant with our GMI caller and gave a concordance rate in the absolute of 84.62 %, one of the SNP was actually a deletion site of 65bp, and the other was wrongly called as Guanine (G base) instead of Adenine (A base).

Finally, we compared our GMI SNP caller with two other SNP callers, MAQ and Samtools, in order to assess the validity of these results (Table 4). Our GMI SNP caller called SNPs in a sensitive way showing the highest number of SNPs among the three callers. The alignment and calling with MAQ called SNPs in a fewer number. SNP calling by Samtools showed a similarity harboring 80% with our GMI SNP caller.

We therefore concluded that a list of common SNPs between GMI caller and Samtools will be a solid base for further analyses. In this regard, we discarded SNPs that were not common between the two SNP calls.



	400										410
Consensus	A	A	A	C	G	T	C	T	C	C	A
Identity											
1. DNA +/- 100bp	A	A	A	C	A SNP7	T	C	T	C	C	A
FWD 2. ES_SNP7_L	A	A	A	C	G	T	C	T	C	C	A
REV 3. ES_SNP7_R	A	A	A	C	G	T	C	T	C	C	A
FWD 4. 1IPS_SNP7_L	A	A	A	C	G	T	C	T	C	C	A
REV 5. 1IPS_SNP7_R	A	A	A	C	G	T	C	T	C	C	A
FWD 6. D0_SNP7_L	A	A	A	C	G	T	C	T	C	C	A
REV 7. D0_SNP7_R	A	A	A	C	G	T	C	T	C	C	A
REV 8. D11_SNP7_R	A	A	A	C	G	T	C	T	C	C	A
FWD 9. D18_SNP7_L	A	A	A	C	G	T	C	T	C	C	A
REV 10. D18_SNP7_R	A	A	A	C	G	T	C	T	C	C	A
FWD 11. 2IPS_SNP7_L	A	A	A	C	G	T	C	T	C	C	A
REV 12. 2IPS_SNP7_R	A	A	A	C	G	T	C	T	C	C	A

Figure 3. Screenshot of SNPs selected randomly and validated by PCR and capillary sequencing. This is a representation of a SNPs validated by capillary sequencing. The reference sequence mm9 build37 is shown on the top called “DNA +/- 100bp” and all SNPs are highlighted in yellow.

Table ４. Summary of SNP genotyping with GMI caller, Samtools and MAQ.

	GMI caller	Samtools		MAQ	
Sample	Total number of SNPs	Total number of SNPs	Common SNPs with GMI caller	Total number of SNPs	Common SNPs with GMI caller
1iPSC	6,430,277	5,583,710	5,107,694	-	-
D0	6,471,732	5,884,030	5,293,149	5,117,579	5,001,875
D11	6,470,187	5,909,966	5,353,262	-	-
D18	6,246,905	5,810,353	5,249,752	-	-
2iPSC	6,348,862	5,577,817	5,042,571	4,913,736	4,733,550

2. Strain differences and homozygosity rate of the five samples

The tremendous amount of SNPs we found exceeds 6 million per sample and is greater than the number of SNPs reported by Keane T.M. *et al.*[18] for a similar mouse strain. The authors reported 4 million SNPs for the strain number 129S1/SvlmJ against C57BL/6J after sequencing in average 71 gigabases of the mouse genome. Whereas, we sequenced a maximum of 150 gigabases of the genome and found over 6 million SNPs. As explained in the material and method section 1, our mouse strain is a combination of two 129 sub-strains and C57BL/6J, which gives a fair comparison with Keane T.M. findings.

From the common list of SNPs called by Samtools and GMI caller, we assessed homozygosity changes during the reprogramming and observed that the ratio of homo-heterozygosity remains stable from 1iPSC to 2iPSC harboring 50% (Figure 4).

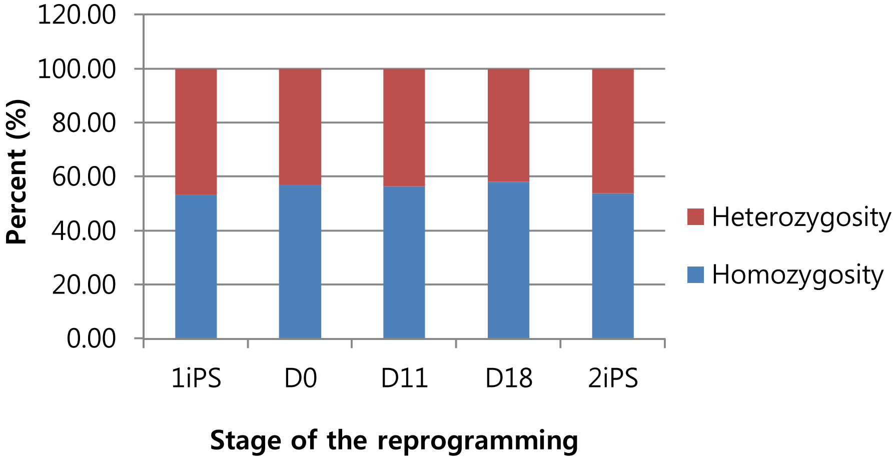


Figure 4. Ratio of homozygosity and heterozygosity throughout the reprogramming from 1iPSC to 2iPSC. Homozygosity is shown in blue and heterozygosity is shown in red, the y-axis represent the fraction of homozygosity (or heterozygosity) versus the total number of SNPs per sample.

2.1. Number of SNPs within gene boundaries and nsSNPs per sample

The number of SNPs found within gene boundaries is stable and ranges from 36 to 38 thousands. We also determined the amount of nsSNPs per sample and identified over 12 thousands nsSNPs for each. As shown in Table 5, the number of nsSNPs is increasing until D11 and then decreasing from D11 to 2iPSC.

Table 5. Summary of SNPs found within gene boundaries and nsSNPs at each step of the reprogramming to pluripotency.

Sample	Number of SNPs	SNPs within gene boundaries	nsSNPs
1iPSC	5,107,694	37,649	13,198
D0	5,293,149	38,300	13,844
D11	5,353,262	38,383	14,022
D18	5,249,752	37,769	13,681
2iPSC	5,042,571	36,151	12,955

2.2. Number of unique SNPs per sample after comparison with 1iPSC

We compared the number of SNPs of D0, D11, D18 and 2iPSC with 1iPSC and extracted the number of unique SNPs for each (Figure 5). We observed that from D11, the midpoint of the timeframe of uncertain cell fate, a light decreasing tendency is present until cells reach the state of 2iPSC.

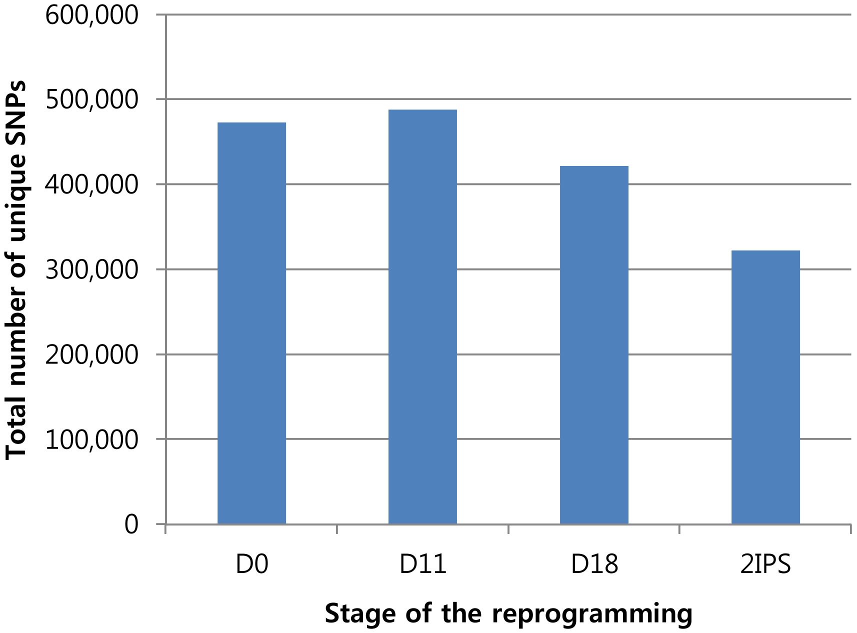


Figure 5. Number of unique SNPs per samples after comparison with 1iPSC. Each sample has been compared one by one with 1iPSC and the number of unique SNPs represents only the number of SNPs that are not shared with 1iPSC.

2.3. Somatic SNPs arising and maintained during the reprogramming to pluripotency

As described in the materials and methods part 3, we listed somatic nsSNPs appearing during the reprogramming in order to understand the development processes occurring at the timeframe of uncertain cell fate. The ratio of somatic sSNPs and nsSNPs increases in favor of nsSNPs from D11 to 2iPSC, corresponding to the end of the timeframe of uncertain cell fate (Table 6, Figure 6).

Table 6. Identification of somatic SNPs maintained during the reprogramming from D0 to 2iPSC.

1iPSC	D0	D11	D18	2iPSC	Somatic SNPs	Somatic nsSNPs
×	✓	✓	✓	✓	171,482	492
×	×	✓	✓	✓	19,560	63
×	×	×	✓	✓	13,244	58
×	×	×	×	✓	48,936	203

N.B.: A cross mark indicates that no SNP was present, whereas a tick mark indicates that SNPs were present at this stage.

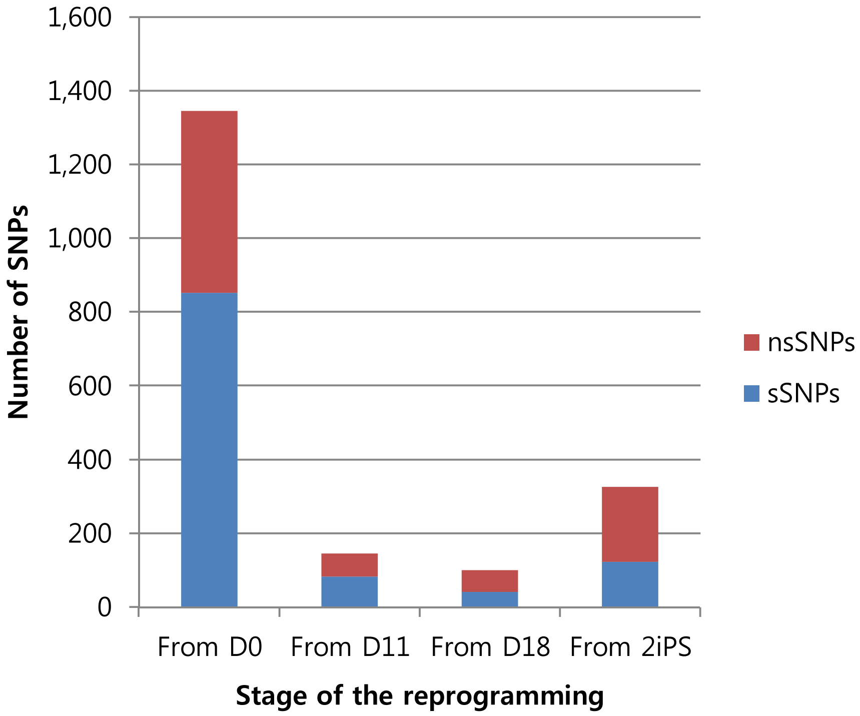


Figure 6. Number of somatic nsSNPs arising from D0 to 2iPSC. The x-axis shows from which day of the development somatic SNPs are arising. The number of nsSNPs displayed accounts only for somatic SNPs maintained through the reprogramming.

2.3.1. Gene Ontology of somatic nsSNPs arising and maintained during the reprogramming

Gene ontology was performed with Ontologizer 2.0 on the three stages spanning the timeframe of uncertain cell fate. Gene ontology of nsSNPs arising before the timeframe of uncertain cell fate didn't show any significant group of gene relevant to the reprogramming. During the midpoint D11, gene ontology shows that tissue development (GO:0009888, p=6.83 x 10^{-3}) is affected by coding mutation in 15 genes. The genes affected include *Gata6*, which is known to be involved in early mouse development[15], and *Myh6*, coding for a heavy polypeptide 6 cardiac muscle alpha.

From D18, cell development (GO:0048468, p=6.51 x10^{-3}) is affected by coding mutations, and regulation of immune response to tumor cells (GO:002837, p=7.94 x 10^{-3}) is also affected by coding mutations in the Mucin 4 gene *Muc4,*and the retinoic acid early transcript gene *Raet1b*.

More interestingly, gene ontology of nsSNPs arising between the end of the timeframe of uncertain cell fate at 2iPSC, shows that the biological process detection of stimulus (GO:009607, p=3.23 x 10^{-4}) is strongly affected by coding mutations through multiple genes of the immune system, notably the histocompatibility gene *H2-K1*, and the gene

Raet1b which is known to be a natural killer cell activation (GO:0030101, p=1.76 x 10^{-3}). Finally epithelial cell proliferation (GO:00050673, p=7.78 x 10^{-4}) is also affected by coding mutations in genes such as *Brca2*, an oncogene well reported as being involved in breast cancer[16]. Coding mutations in the *Wnt* gene seem to affect muscle cell differentiation (GO:0042692, p=9.38 x 10^{-3}) and regulation of cell-cell adhesion (GO:0022407, p=8.64 x 10^{-3}), suggesting that the Wnt signaling pathway is also potentially affected. In general, 11% of mutations arising between D18 and 2iPSC (Figure 7) affect cell differentiation, 26% affect gene expression (GO:0010467, p=9.05 x 10^{-3}), 21% affect the immune system and 16% affect the response to stimulus (GO:0009607, p=3.22 x 10^{-4}).

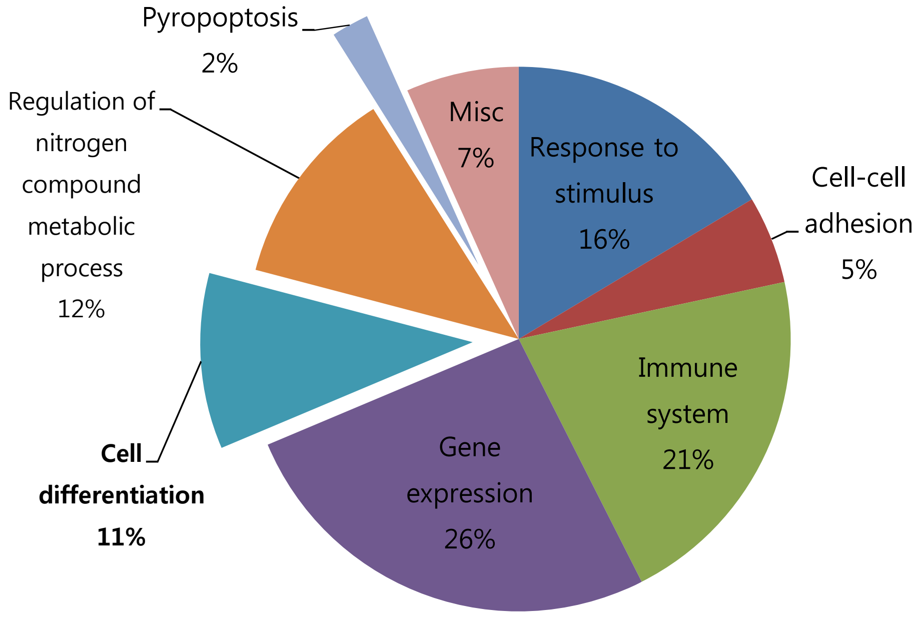


Figure 7. Representation of GO terms found by gene ontology for somatic nsSNPs arising between D18 and 2iPSC.

Table 7. Gene ontology summary of nsSNPs appearing at D11.

GO term ID	p-value	GO term	Genes
GO:0071346	4.35×10^{-3}	Cellular response to interferon-gamma	Gbp1, Gbp10, H2-Ab1
GO:0009888	6.83×10^{-3}	Tissue development	Col11a2, Creb3l2, Crhr1, Ddr1, Dll1, Etv4, Fgf4, Gata6, H2-Ab1, Hspg2, Myh6, Ovol2, Pdzrn3, Ryr2, Tapbp
GO:0046620	6.98×10^{-3}	Regulation of organ growth	Gata6, Myh6
GO:0060420	7.37×10^{-3}	Regulation of heart growth	Gata6, Myh6
GO:0034341	7.78×10^{-3}	Response to interferon-drama	Gbp1, Gbp10, H2-Ab1
GO:0002682	9.48×10^{-3}	Regulation of immune system process	C1qc, Crhr1, Dll1, H2-Ab1, H2-Bl, Itpkb, Muc4, Spta1, Tlr9

Table 8 . Gene ontology summary of nsSNPs appearing at D18.

GO term ID	p-value	GO term	Genes
GO:1901700	1.39×10^{-3}	Response to oxygen containing compound	EU599041, Gbp10, Nanog, Raet1a, Raet1b, Ren1, Ssh1
GO:0008154	3.96×10^{-3}	Actin polymerization or depolymerization	Arpc1b, Map3k1, Ssh1
GO:0009607	5.82×10^{-3}	Response to biotic stimulus	Gbp10, Ifna5, Ifna9, Itln1, Muc4, Raet1a, Raet1b
GO:0048468	6.51×10^{-3}	Cell development	Cdh11, Myh6, Nanog, Pdzrn3, Ren1, Sox11, Ssh1, Tdrd6, Tiam1, Tnn
GO:0002831	7.07×10^{-3}	Regulation of response to biotic stimulus	Muc4, Raet1b
GO:0051100	7.24×10^{-3}	Negative regulation of binding	Mdfi, Sox11
GO:0002837	7.94×10^{-3}	Regulation of immune response to tumor cell	Muc4, Raet1b
GO:0030509	8.13×10^{-3}	BMP signaling pathway	Myh6, Nanog, Sox11
GO:0007257	8.33×10^{-3}	Activation of JUN kinase activity	Map3k1, Mdfi
GO:0031341	8.66×10^{-3}	Regulation of cell killing	Klrb1c, Muc4, Raet1b
GO:0071216	8.81×10^{-3}	Cellular response to biotic stimulus	Gbp10, Raet1a, Raet1b
GO:0002717	9.52×10^{-3}	Positive regulation of natural killer cell mediated immunity	Klrb1c, Raet1b

Table 9. Gene ontology summary of nsSNPs appearing at 2iPSC.

GO term ID	p-value	GO term	Genes
GO:0009607	3.22 x 10^{-4}	Response to biotic stimulus	Gbp1, H2-K1, Ifna2, Ifna5, Ifna9, Itln1, Muc4, Naip2, Naip5, Naip6, Nlrp1a, Plscr1, Raet1a, Raet1b
GO:0030101	1.76 x 10^{-3}	Natural killer cell activation	Elf4, H60b, H60c, Ifna9, Raet1b
GO:0022408	1.83 x 10^{-3}	Negative regulation of cell-cell adhesion	Gm9573, Muc4, Wnt1
GO:0009116	1.99 x 10^{-3}	Nucleoside metabolic process	Atp2b2, Mtor, Myh14, Nudt9, Plscr1, Smarca4
GO:0006952	2.15 x 10^{-3}	Defense response	Darc, Gbp1, H2-K1, Herc6, Ifna2, Ifna5, Ifna9, Mrgpra3, Naip2, Naip5, Naip6, Nlrp1a, Plscr1, Raet1a, Raet1b
GO:0070269	2.79 x 10^{-3}	Pyroptosis	Naip2, Naip5, Naip6
GO:0051173	3.72 x 10^{-3}	Positive regulation of nitrogen compound metabolic process	Brca2, Creb3l2, Elf4, Evx1, Fhod1, Foxd4, Mtor, Osr2, Plscr1, Prpf19, Raet1b, Smarca4, Tbx3, Wnt1, Ybx1, Zfp384
GO:0071391	4.55 x 10^{-3}	Cellular response to estrogen stimulus	Naip1, Naip2, Naip6
GO:0045807	4.81 x 10^{-3}	Positive regulation of endocytosis	Cd63, Sirpa, Sirpb1a
GO:0042742	4.99 x 10^{-3}	Defense response to bacterium	Gbp1, H2-K1, Naip2, Naip5, Naip6, Nlrp1a, Raet1a, Raet1b
GO:0009581	5.72 x 10^{-3}	Detection of external stimulus	Atp2b2, Cacna1f, Naip2, Naip5, Naip6
GO:0050673	7.79E-03	Epithelial cell proliferation	Brca2, Col8a2, Mtor, Osr2, Thap1, Wdr77
GO:0022407	8.64E-03	Regulation of cell-cell adhesion	Gm9573, L1cam, Muc4, Wnt1
GO:0010467	9.05E-03	Gene expression	Brca2, Cdk4, Cdk8, Creb3l2, Dach2, Eef1g, Elf4, Evx1, Fhod1, Foxd4, Grhl1, Gtf3a, Larp7, Mbd1, Mrpl49, Mtor, Mybl2, Osr2, Plscr1, Prickle1, Prpf19, Rcor2, Ring1, Rps2, Rps24, Smarca4, Tbx3, Thap1, Timeless, Trmt6, Wbp11, Wdr77, Wnt1, Ybx1, Zfp384
GO:0042692	9.38E-03	Muscle cell differentiation	Mtor, Ryr1, Smarca4, Tbx3, Ttn, Utrn, Wnt1, Ybx1

2.4. SNPs found within miRNA of 1iPSC and 2iPSC

It was reported that reprogramming induced with miRNA has been successful[17]. We therefore selected only SNPs within boundaries of genes coding for miRNA and questioned if some of these SNPs could influence the fate of cell reprogramming to pluripotency. As seen in Table 7, our SNP caller identified a total of 221 SNPs within miRNA genes including 198 *novel* SNPs. We noticed a number of SNPs in the gene coding for *Mirg*, a micro RNA recently identified as being involved in the full development of iPSC[18].

Table 10. Summary of SNPs found in gene coding for miRNA.

Chromosome	Gene	Total Number of SNPs	Gene Length (bp)
2	*Mir129-2*	1	90
2	*Mir296*	2	79
8	*Mir1186*	1	122
8	*Mir1969*	1	94
8	*Mir24-2*	1	107
9	*Mir1899*	1	96
9	*Mir3471-1*	3	123,979
9	*Mir184*	1	69
12	*Mir680-3*	1	87
12	*Mir673*	1	91
12	*Mir882*	1	77
12	*Mir1197*	1	120
12	*Mir654*	1	84
12	*Mirg*	120	14,477
14	*Mir1971*	1	106
15	*Mir297-1*	7	70,622
16	*Mir1946a*	1	134
17	*Mir692-1*	1	109
19	*Mir3086*	4	87

3. Indels found within the five samples

Our GMI caller has identified more than 800 thousand of indels against mm9 (UCSC, build37). Subsequent annotation of indels shows that only 1% of them are within CDS. Among indels within gene boundaries, around 1% causes a frameshit changes which represent about 100 indels among 800,000 initially found. An increase of indels causing a frameshift is seen from D0, reaching a maximum at D11 and continuously decreasing until 2iPSC. Unlike our SNP data the timeframe of uncertain cell fate did not show a strong increase of indels with a clear demarcation from D11 to 2iPSC. However we can still observe a tendency to a higher number of variants inducing a frameshit at D11 (Figure 8, Table 8).

Table 11. Summary of the number of indels found within the five iPSC.

Sample	Total number of indels	Indels within gene boundaries	Indels causing frameshifts
1iPSC	856,585	9,225	133
D0	835,030	9,107	142
D11	850,807	9,304	145
D18	799,105	8,743	135
2iPSC	830,910	8,889	129

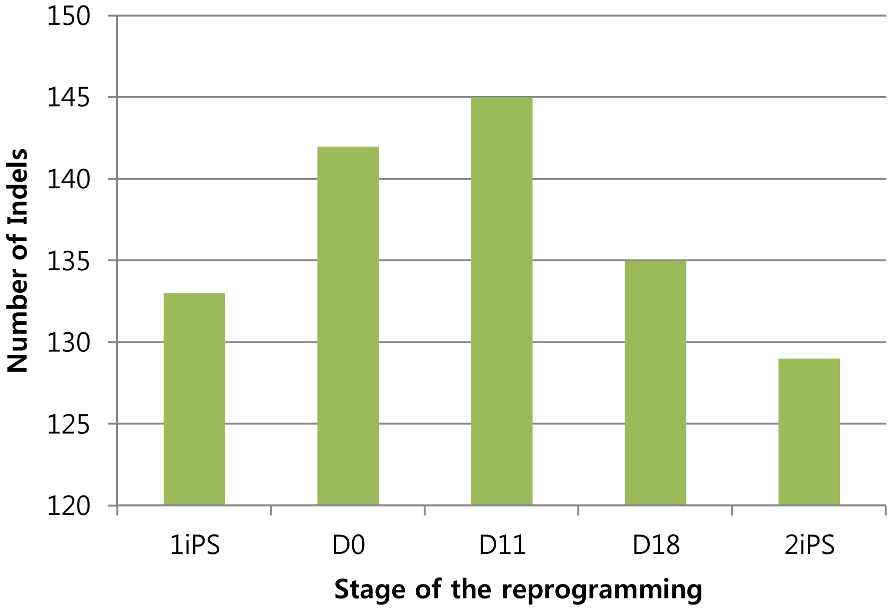


Figure 8. Representation of the number of indels causing frameshifts in the genome.

DISCUSSION

One of our main goals was to validate the theory highlighted by Hussein S.M. *et al.*, stating that the highest number of mutations was observed in iPSC during the midpoint of the reprogramming, and 52% these variants were iPSC specific[19]. We then explored pathways potentially involved in the reprogramming process. Hussein S.M. *et al.* based their work on Affymetrix SNP array, which offers less resolution compared to the WGS. We were able to perform a WGS of 5 samples at high throughput and genotype SNPs and indels.

The first obstacle of this study was to determine whether the high volume of SNPs identified were true or not. To identify if such a high volume of variants was due to the strain difference, we compared our results with those of a similar strain. A study published in 2011 by Keane T. *et al.* reported that the mouse strain 129S1/SvlmJ was carrying 4,458,004 SNPs identified against the reference strain C57BL/6J[20]. Our mouse strain is a hybrid of the two 129 sub-strains 129X1/SvJ and 129S1/Sv-Oca2+ Tyr+ KitlSl-J and the strain C57BL/6J and this mixed strain is similar to the one studied by Keane T. *et al.* Thus, after validation by capillary sequencing, the true positive rate was of at least 84.62%. Finally, two additional SNP calls with MAQ and Samtools showed a similar number of variants with at least

80% similarity with our GMI SNP call. The homozygosity and heterozygosity rate was around 50% and remained stable during the reprogramming.

Among the five samples, a brief comparison of the number nsSNPs (Table 5) and unique SNPs (Figure 5) reveals that a maximum is reached at D11, the midpoint of the timeframe of uncertain cell fate. This tendency to a higher number of variants during the reprogramming before observing decreased trend at 2iPSC has been reported by Hussein S.M. *et al.*[19] with copy number variants. However, since our sequencing summary shows a similar trend for the coverage depth, it is then difficult to conclude that this Gaussian curve pattern shown throughout the samples is only due to the reprogramming process.

To confirm the theory elaborated by Ji J. *et al.*[21] stating that iPSC were carrying an elevated coding mutation rate we questioned the number of somatic nsSNPs. Our results show that the proportion of nsSNPs is increasing from D11 to 2iPSC. Subsequent gene ontology revealed that no group of gene was significantly involved at the beginning and midpoint of the timeframe of uncertain cell fate, even though more than 400 coding mutations were identified at this point. At D18, gene ontology showed that cell development and the regulation of immune response were affected by coding mutations (Table 9). Between the end of the timeframe of uncertain cell fate and the end of the reprogramming, gene ontology showed that coding mutations significantly affect group of genes such at pyropoptosis, cell differentiation and cell-cell adhesion. Genes carrying coding mutation include the oncogene *Brca2* and the *Wnt* gene, directly involved in the Wnt pathway (Table 10).

As micro RNAs (miRNA) have recently been reported to be involved in the reprogramming process of iPSC[9] we queried the number of SNPs within miRNA. We found SNPs within boundaries of 28 miRNA clusters. Notably the cluster of *Mirg* contains 138 SNPs of which 123 are *novel*. Others miRNAs have a gene length ranging from 70bp to 100bp and carried 1 to 7 variants. The great amount of SNPs found in *Mirg* could be due to its size. Indeed, the gene *Mirg* has a length of 14,477bp. However two others miRNAs, miR1971 and miR3471-1, having size respectively of 70,622bp and 12,979bp did not carry more than 7 variants, suggesting that the great amount of variants found in *Mirg* should be further characterized. *Mirg* is a non-coding RNA exhibiting sustained expression throughout mouse embryogenesis from Embryo day 8.5 to Embryo day 18.5, with a maximum of expression levels at Embryo day 15.5. The first author announcing the importance of *Mirg* in the reprogramming to pluripotency underlined that the correct expression of the maternal copy of *Mirg* is a marker for the full developmental potential of iPSC. Transcriptional silencing of the maternal copy of *Mirg* resulted in failure to generate all-iPSC mice, which indicates that *Mirg* may have a significant contribution to mouse embryogenic development[22]. The samples we studied were yet at the embryonic stage of development but at the step of reprogramming to pluripotency. Therefore, it is possible that *Mirg* was silenced in order to pause cell differentiation process and let the reprogramming take place. Despite its large size, this large amount of SNPs within boundaries of the *Mirg* cluster suggests that transcriptional activity is being either enhanced or repressed at this site.

Indels potentially causing a change in protein structure are indels causing frameshits. As shown with the amount of nsSNPs and unique SNPs per sample and as mentioned by Hussein S.M. *et al.*, an increasing tendency of variants with a Gaussian shape is present in the window of the timeframe of uncertain cell fate. Once again, according to the fact that the coverage depth also follows the same curve, it is hard to distinguish if this pattern is only due to the reprogramming or to the sequencing throughput. Therefore these data should be interpreted with care. If true, it is interesting to note that the authors also suggested that an increased number of variants is necessary for the reprogramming to pluripotency, however negative selection of highly mutated cells might during the timeframe of uncertain cell fate[19], which would explain why only a fraction of variations persist throughout the reprogramming.

CONCLUSION

Our results show that a number of variations occurring during the timeframe of uncertain cell fate were confirmed by two different algorithms. We were able to observe a light decreased tendency of variants from the midpoint of the timeframe of uncertain cell fate at D11 until the end of the reprogramming at 2iPSC. Thus, our results give us insights on mechanisms involved during the reprogramming to pluripotency. Gene ontology analysis has shown evidences that coding mutations affected cell differentiation, gene expression and the immune system at the end of the reprogramming. We therefore encourage further study to characterize in detail the role of variants in the reprogramming.

ACKNOWLEDGEMENT

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable knowledge in the preparation and completion of this study.

First and foremost, my utmost gratitude to Prof. Seo Jeong Sun, as my advisor and chairman of Macrogen Inc., I will never forget that Prof. Seo Jung Sun has been my guidance and leader as I hurdle all the obstacles in the completion of this research work.

Dr. Shin J.Y., manager of the iPS cell project at the genome medical institute (GMI), Seoul National University, College of Medicine, who had kind concern and consideration regarding my academic requirements and the achievement of this study.

Dr. Lee SB, for his patience and steadfast encouragement to complete this study in *bonne et due forme*

Dr. Hong DW and Park SS, staff member of PSOMA and GMI for their guidance and support through the establishment of algorithms for the analysis of our samples.

The staff of GMI and members of Professor Seo's laboratory for all the help and devotion.

And finally, a very special thanks to Dr. Nagy and Dr. Puti Mira, Stem Research Center in Toronto Ontario, for handling us the five samples of this study and for always being supportive during my stay in Korea.

국문 초록

서론: 유도만능줄기세포(iPSC)의 리프로그래밍 성공에도 불구하고 유도만능줄기세포의 전유전체서열은 한번도 고처리량으로 해석된 적이 없다. 유도만능줄기세포는 임상 응용에서 치료제로서 배아줄기세포를 대체할 능력을 가지고 있으므로 리프로그래밍 과정에 발생하는 유전체변화를 이해하는 것이 중요하다. Nagy A. *et al.* 의 실험 등에서 보고된 것과 같이, 유도만능줄기세포의 실험결과는 실험 5 일째부터 15 일째까지의 기간 동안에는 불확실하며, 16 일 이후에는 전분화능으로의 리프로그래밍이 반드시 일어났다. 리프로그래밍과 유전체 변이 발생과정에서 일어난 메커니즘을 평가하기 위하여 우리는 생쥐 섬유아세포로부터 생성된 유도만능줄기세포에서 발생하는 구조 변형에 대하여 연구하였다. 이 연구는 timeframe of uncertain cell fate, 즉 실험 5 일째부터 15 일째까지의 기간을 포함, 전후에도 이루어졌다. **방법:** 우리는 차세대 시퀀서 Illumina HighSeq 2000 을 사용하여 평균 5 개의 샘플에서 35.77X 의 높은 범위로 전유전체서열을 생성하였다. 5 개의 샘플은 초기 유도만능줄기세포(1iPSC) ,리프로그래밍의 각 3 단계(실험 시작 당일,실험 시작 후 11 일, 18 일)의 샘플들 그리고 2 차 유도만능줄기세포이다. 우리는 또한 자사의 GMI 호출기를 사용하여 단일염기 다형성(SNPs)과 단기 삽입 제거 유전형분석을 수행하였다. 이 방법을 두 개의 다른 알고리즘인 Samtools 와 MAQ 과 비교했다. 그 다음의 gene ontology 는 Ontologizer2.0 을 사용하여 수행했다. **결과:** 우리는 매 샘플 당 5 백만개 이상의 단일염기 다형성 (SNP)들을 확인하였으며, 초기

유도만능줄기세포와 비교하였을 때 각 샘플이 가지고 있는 non-synonymous SNPs, 즉 고유 SNP 의 수와 insertion-deletion 의 수가 11 일째에 최고에 다다른 후 2 차 유도만능줄기세포 등장 전까지 줄어드는 것으로 나타났다. Ji J. *et al.* 의 연구등에서 보고되었듯이, 전분화능으로 리프로그래밍이 되는 세포들에서 11 일째 날로부터 2 차 유도만능줄기세포 전까지 더 많은 수의 체세포 암호화 돌연변이를 발견할 수 있었다. 리프로그래밍 후반에 발생한 체세포 nsSNPs 의 gene ontology 를 통하여, 유전자 발현(p=9.05 x 10^{-3}), 세포분화(p=9.38 x 10^{-3}), 그리고 세포증식(p=7.79 x 10^{-3})과 같은 생물학적 과정은 coding mutations 이들에 영향을 받는다는 것을 밝혔다. 이 결과들을 통하여 알 수 있는 것은, 위와 같은 생물학적 과정을 겪은 유전자들이 전분화능으로의 리프로그래밍 과정에 영향을 미친다는 것이다. **결론:** Timeframe of uncertain cell fate 는 리프로그래밍 과정에서 중요한 의미를 가지는 제한적 시간이다. Ji J. *et al.*의 실험등에서 이미 보고되었듯이, 이번 실험 또한 리프로그래밍 동안 coding mutation rate 가 증가함을 확인할 수 있었다. 마지막으로 gene ontology 는 coding mutations 이 리프로그래밍 후반에 세포 분화, 면역체계, 그리고 세포 발현등과 같은 주요 생물학적 과정에 영향을 끼친다는 것을 밝혔다. 그러므로, timeframe of uncertain cell fate 동안 발생하는 메커니즘들을 이해하기 위하여 위와 같은 변수들에 대한 연구를 더 많이 권장한다.

주요어 : 유도만능줄기세포, 줄기세포, 리프로그래밍, 4 야마나카 전사요소, 전유전체서열, 단일염기 다형성, 삽입 또는 결실

학 번 : 2011-23019

REFERENCES

1. Baker M. Adult cells reprogrammed to pluripotency, without tumors. Nature Reports Stem Cells. 2007; doi:10.1038/stemcells.2007.124

2. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006; 126(4): 663-76.

3. Li Z, Rana TM. Using microRNAs to enhance the generation of induced pluripotent stem cells. Curr Protoc Stem Cell Biol. 2012; Chapter 4: Unit 4A.4.

4. Maekawa M, Yamaguchi K, Nakamura T, Shibukawa R, Kodanaka I, Ichisaka T, Kawamura Y, Mochizuki H, Goshima N, Yamanaka S. Direct reprogramming of somatic cells is promoted by maternal transcription factor Glis1. Nature. 2011; 474(7350): 225-9.

5. Takahashi K, Tanabe K, Ohnuki M, *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007; 131(5): 861-72.

6. Nagy A, Nagy K. The mysteries of induced pluripotency: where will they lead? Nat Methods. 2010; 7(1): 22-4.

7. Ju YS, Kim JI, Kim S, Hong D, Park H, Shin JY, Lee S, Lee WC, Yu SB, Park SS, *et al.* Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. Nat Genet. 2011; 43(8): 745–752.

8. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, *et al.* A highly annotated whole-genome sequence of a Korean individual. Nature. 2009; 460(7258): 1011–5.

9. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26(7): 873-81.

10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25(14): 1754-60.

11. McKenna A, Hanna M, Banks E, *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20(9):1297–1303.

12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16): 2078-9.

13. Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics. 2008; 24(14): 1650-1.

14. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29(1): 308-11.

15. Lavial F, *et al.* Bml facilitates primitive endoderm formation by stabilizing Gata6 during early mouse development. Genes Dev. 2012;

26(13): 1445-58.

16. Bosviel R, Durif J, Guo J, Mebrek M, Kwiatkowski F, Bignon YJ, Bernard-Gallon DJ. BRCA2 Promoter Hypermethylation in Sporadic Breast Cancer. OMICS. 2012; 16(12): 707-10.

17. Underbayev C, Kasar S, Yuan Y, Raveche E. MicroRNAs and induced pluripotent stem cells for human disease mouse modeling. J Biomed Biotechnol. 2012; 2012: 758169.

18. Han Z, He H, Zhang F, Huang Z, Liu Z, Jiang H, Wu Q. Spatiotemporal expression pattern of Mirg, an imprinted non-coding gene, during mouse embryogenesis. J Mol Histol. 2012; 43(1): 1-8.

19. Hussein SM, *et al.* Copy number variation and selection during reprogramming to pluripotency. Nature. 2011; 471(7336): 58-62.

20. Keane TM, *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011; 477(7364): 289-94.

21. Ji J, Ng SH, Sharma V, Neculai D, Hussein S, Sam M, Trinh Q, Church GM, McPherson JD, Nagy A, Batada NN. Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. Stem Cells. 2012; 30(3): 435-40.

22. Stadtfeld M, Apostolou E, Akutsu H, Fukuda A, Follett P, Natesan S, Kono T, Shioda T, Hochedlinger K. Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells.

Nature. 2010; 465(7295): 175–181.