Genomics 104 (2014) 121–127

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

High-throughput single nucleotide variant discovery in E14 mouse embryonic stem cells provides a new reference genome assembly

Danny Incarnato [a,b], Anna Krepelova [a], Francesco Neri [a,*]

[a] Human Genetics Foundation (HuGeF), via Nizza 52, 10126 Torino, Italy
[b] Dipartimento di Biotecnologie, Chimica e Farmacia, Università degli Studi di Siena, Via Fiorentina 1, 53100 Siena, Italy

ARTICLE INFO

Article history:
Received 25 March 2014
Accepted 27 June 2014
Available online 5 July 2014

Keywords:
SNV
Genotyping
Sequencing
ESC
E14
Genome reference

ABSTRACT

Mouse E14 embryonic stem cells (ESCs) are a well-characterized and widespread used ESC line, often employed for genome-wide studies involving next generation sequencing analysis. More than 2×10^9 sequences made on Illumina platform derived from the genome of E14 ESCs were used to build a database of about 2.7×10^6 single nucleotide variants (SNVs). The identified variants are enriched in intergenic regions, but several thousands reside in gene exons and regulatory regions, such as promoters, enhancers, splicing sites and untranslated regions of RNA, thus indicating high probability of an important functional impact on the molecular biology of these cells. We created a new E14 genome assembly reference that increases the number of mapped reads of about 5%. We performed a Reduced Representation Bisulfite Sequencing on E14 ESCs and we obtained an increase of about 120,000 called CpGs and avoided about 20,000 wrong CpG calls with respect to the mm9 genome reference.

1. Introduction

In the last two decades, the world's interest has focused on the study of embryonic stem cells since they represent an important tool for the understanding of developmental processes, stemness homeostasis and self-renewal and are essential for gene targeting in mouse models. Thousands of papers have been published describing the genomic and proteomic regulations that rule this particular cell line [1–3].

E14 embryonic stem cells (ESCs) have been isolated many years ago from mouse blastocyst derived from a strain of 129/Ola and still are one of the most widespread ESC lines used in laboratory. These cells can be maintained undifferentiated in culture, and they can be differentiated with the appropriate stimuli into each lineage, or injected into the blastocyst to generate a chimeric mouse [3–6]. More recent studies have revealed that the understanding of the ESC's biology is very important also for the comprehension of the cancer development and progression [7–12].

Genomic, epigenetic, and transcriptomic profiles of E14 ESCs have been largely addressed and deeply studied. In the last years, thousands of papers and hundreds of next generation sequencing data on E14 embryonic stem cells have been published and deposited on public databases such as GEO Datasets or ENCODE [2,13–22]. Mapping of these data is usually performed on the mm9 genome assembly (MGSCv37 in NCBI), even if this assembly derives from the C57BL/6J mouse strain. We performed the genotyping of the E14 mouse ESC line, and discovered about 2.7 millions of single nucleotide variants (SNVs) with respect to the MGSCv37 assembly. From this E14 genotype we assembled a new E14 genome reference to provide a tool for all the scientific community working with next generation data produced in this cell line. Indeed, by using this new E14 reference assembly we observed an improvement in the mapping efficiency of sequencing data. Moreover, when analyzing bisulfite-sequencing data we avoided many false methylation status calls. Interestingly, many of the identified SNVs reside within exons or regulatory regions and 28 SNVs insert a new improper stop codon. SNVs in such regions, may lead to alterations of the biology of this cell line, thus making our SNV dataset useful also for all the scientists belonging to the ESC research area that use E14 cell line or its derivatives.

2. Material and methods

2.1. Illumina sequencing

For SNV calling we pooled all genome sequencing datasets for E14 produced by our lab using Illumina HiScan SQ platform. Libraries were prepared using the Illumina DNA Sample Prep Kit according to the manufacturer's instructions. Prior to mapping, read quality was estimated using FastQC tool v0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Nucleotide positions with a quality score behind 30 (Phred33 scale) were trimmed using the *fastq_quality_filter* tool

Abbreviations: SNVs, single nucleotide variants; ESCs, embryonic stem cells; Gb, gigabyte; mm9, *Mus musculus* genome assembly reference number 9; ChIP-Seq/MeDIP-Seq, chromatin/methylated DNA immunoprecipitation sequencing; m/nc/mi/LincRNA, messenger/non-coding/micro/long intergenic non-coding RNA; WCE, whole cell extract.

* Corresponding author.
E-mail address: francesco.neri@hugef-torino.org (F. Neri).

http://dx.doi.org/10.1016/j.ygeno.2014.06.007
0888-7543/

from the FASTX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) using the following parameters: −Q 33 −q 30 −p 90.

After low-quality position trimming, reads in which sequencing continued through the 3′ adapter sequence were subjected to adapter clipping using the *fastx_clipper* tool from the FASTX Toolkit using the following parameters: −a TGGAATTCTCGGGTGCCAAGG −l 35 −M 10 −Q 33.

2.2. SNV calling pipeline and E14 genome reference assembly

Reads were mapped using Bowtie [23] to the mm9 genome assembly (MGSCv37 in NCBI) (http://genome.ucsc.edu/cgi-bin/hgGateway?clade=mammal&org=Mouse&db=mm9) allowing up to two mismatches and keeping only uniquely mapped reads using the following parameters: −n 2 −m 1. SAM output file was converted to BAM format using SAMtools v0.1.19 (http://samtools.sourceforge.net/), and file was reference sorted [24]. To remove possible false positive calls due to PCR over-amplification artifacts, prior to variants calling, reads with the same mapping positions were collapsed into one using the *rmdup* tool from SAMtools. Variant calling was performed using the *mpileup* tool from SAMtools using the following command: *samtools mpileup −uf mm9.fa file.bam 2 > stderr.txt | bcftools view −vcg − > snp.vcf.*

Next, we used a customized version of VCFtools v0.1.11 (http://vcftools.sourceforge.net/, customized version available upon request) to select only SNVs with coverage ≥10 and a frequency ≥0.5 using the following parameters: *varFilter −d 10 −a 0.5*. Moreover, using custom Perl scripts (available upon request) we discarded sites with more than one variant call at the same place. Then, using the GATK v2.7-4 (http://www.broadinstitute.org/gatk/) *FastaAlternateReferenceMaker* function with the default parameters, we produced a reference E14 assembly starting from the mm9 genome assembly.

2.3. Reduced Representation Bisulfite Sequencing (RRBS)

1 μg of mESC genomic DNA was digested for 4 h at 37 °C with 200 U of MspI restriction endonuclease (NEB). Digested DNA was then end repaired, dA-tailed, and ligated to methylated adapters, using the Illumina TruSeq DNA Sample Prep Kit, following the manufacturer's instructions. Adapter-ligated DNA was loaded on an EGel Size select 2% agarose pre-cast gel (Invitrogen), and a fraction corresponding to fragments ranging from 180 bp to 350 bp was recovered. Purified DNA was then subjected to bisulfite conversion using the EpiTect Bisulfite Kit (Qiagen). Bisulfite-converted DNA was finally enriched by 15 cycles of PCR using Pfu Turbo Cx HotStart Taq (Agilent). The bisulfite conversion rate was >99.5% calculated on the number of converted cytosines in CpT context.

2.4. Cell culture

The ES cell line E14 was derived from the inbred mouse strain 129/Ola in 1985 by Dr. Martin Hooper in Edinburgh, Scotland [4]. Cells were further expanded in several different laboratories. Thus, the cells used for this study were cultured for at least ten passages on feeder-free gelatin-coated plates in DMEM high glucose medium (Invitrogen) supplemented with 15% FBS (Millipore), 0.1 mM nonessential amino acids (Invitrogen), 1 mM sodium pyruvate (Invitrogen), 0.1 mM 2-mercaptoethanol, 1500 U/ml LIF (Millipore), 25 U of penicillin/ml, and 25 μg of streptomycin/ml.

2.5. Third-party datasets

Datasets used for comparison of E14 SNVs among different laboratories were obtained from the Gene Expression Omnibus for Chen's (GSE11431) and Helin's (GSE24843) laboratories. For the comparison with the 129/Ola strain, we downloaded ~5×10^8 sequencing reads from the Wellcome Trust Sanger Institute Mouse Genomes Project (http://www.sanger.ac.uk/resources/mouse/genomes/). All data was analyzed as stated in Section 2.2 of the Material and Methods section.

3. Results

3.1. SNV distribution

About 115×10^9 bases were used to cover almost 90% of the total mouse genome with an average coverage of 51× (Fig. 1A). Coverage

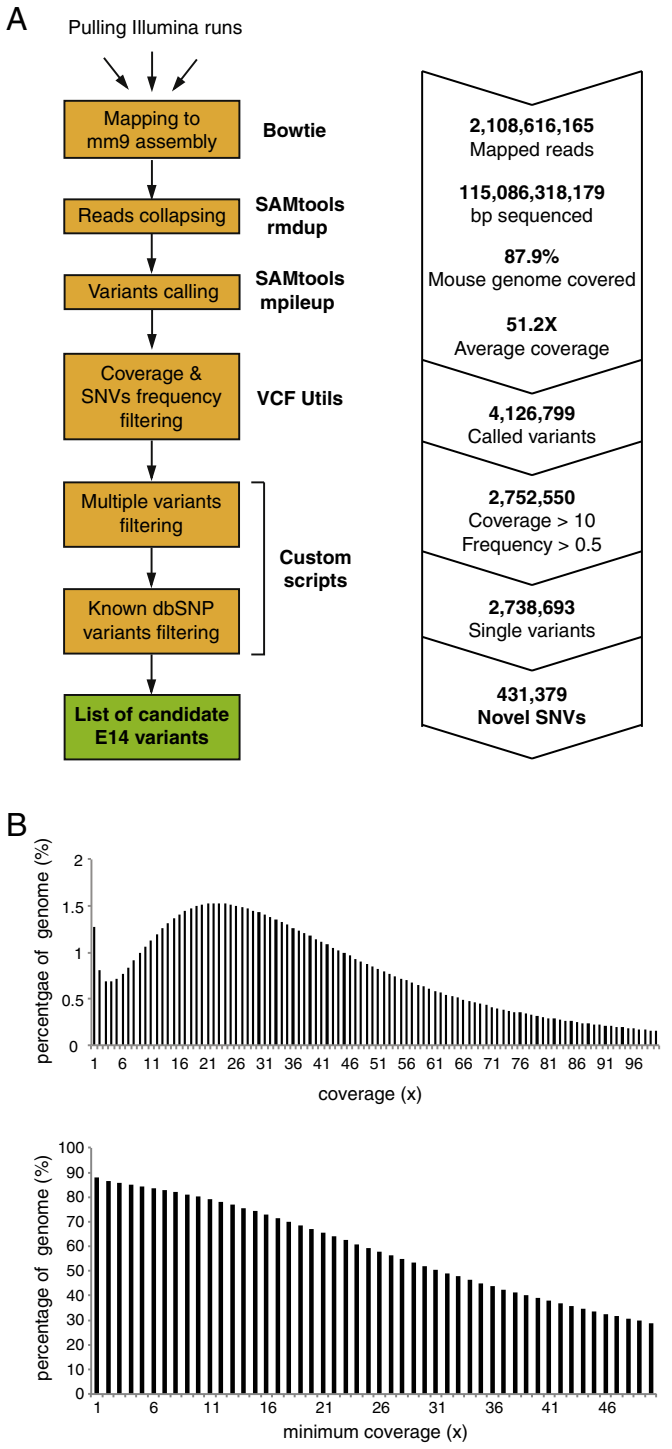


Fig. 1. A) Scheme of the pipeline used in this study to identify SNVs in E14 ESC genome. B) Coverage distribution histogram. C) Histogram showing the percentage of genome covered at the indicated coverage on X-axis.

distribution and total coverage graphs show that the majority of the genome was covered at least 20× (Figs. 1B–C). SNVs were called and filtered using SAMtools, VCF utils and custom scripts and finally a list of 2,738,693 SNVs was obtained. Of the discovered variants, 431,379 (15.7%) represent novel SNVs (Fig. 2A). The variants are equally distributed on the four nucleotides (A,T,C,G) with a higher chance of observing purine-to-purine or pyrimidine-to-pyrimidine variations (Figs. 2B–C). As expected, the called variants are enriched in intergenic regions with respect to promoter or intragenic regions (Figs. 2D–E). No significant enrichment was observed on active (H3K4me3 −/H3K4me1 +/H3K27ac +) or poised (H3K4me3 −/H3K4me1 +/H3K27ac −) enhancers [25–27] (Fig. 2F). Moreover, the intragenic variants are enriched in introns with respect to the exons (Fig. 2G).

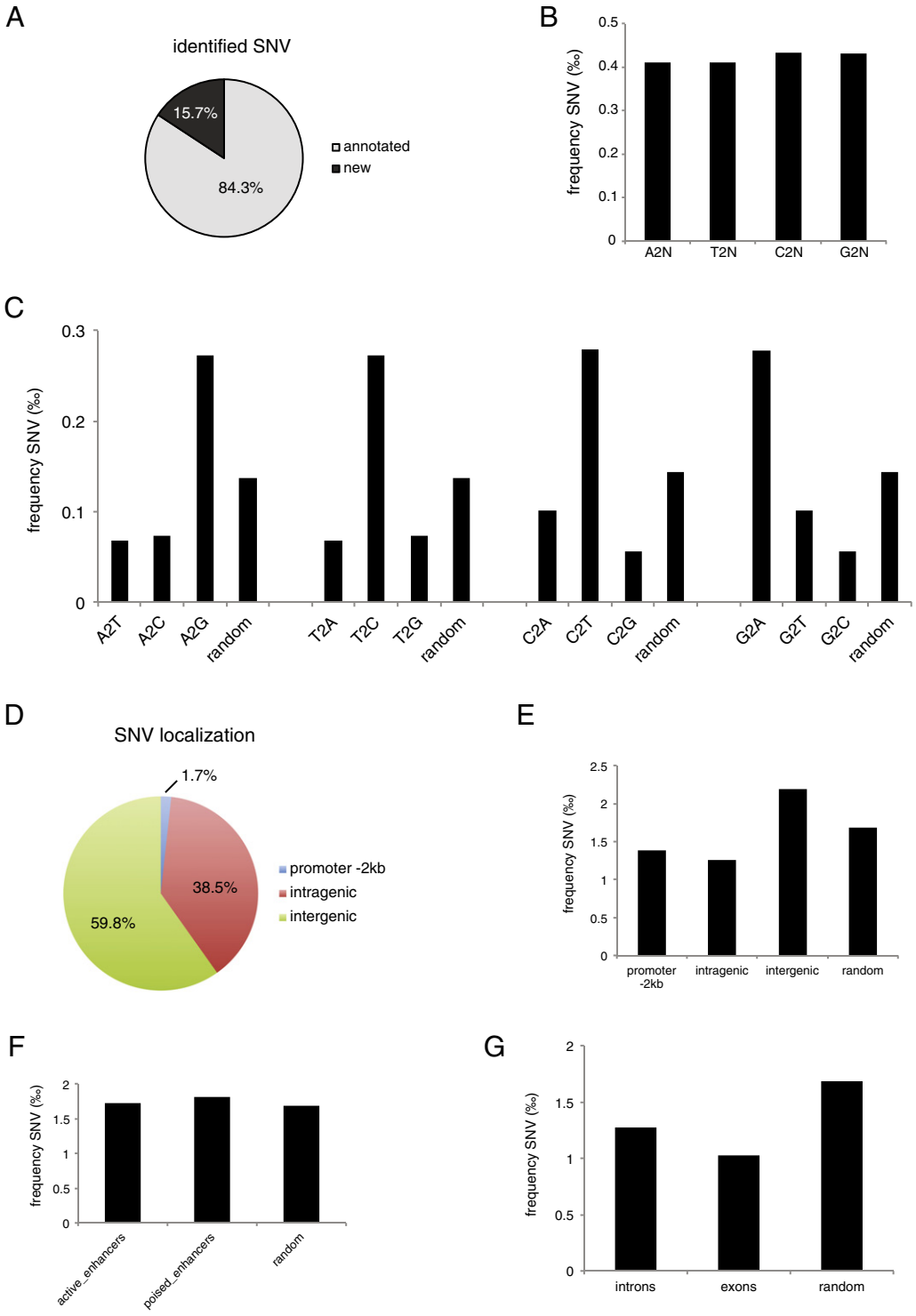


Fig. 2. A) Venn diagram showing the percentage of known and novel SNVs identified. The known SNV dataset (SNPs 128) was downloaded from UCSC (http://genome-euro.ucsc.edu/cgi-bin/hgTables?command=start). B–C) Histograms showing the four bases of exchange frequency (in ‰) versus the other bases. D) Venn diagram with the genomic localization of the SNVs identified in E14 genome. E–F–G) The bar graphs represent the SNV frequency (in ‰) in the indicated genomic regions. Random frequency was calculated using several random datasets of equal size to the SNV datasets generated using randomBed (bedtools random Version: v2.17.0).

3.2. Dataset validation

To validate our dataset, raw data for ChIP-Seq experiments generated by Helin's (1st dataset, GSE19365-GSE24843-GSE37930) and Chen's (2nd dataset, GSE11431) laboratories in E14 mouse embryonic stem cells were obtained from GEO Datasets database. We analyzed these validation datasets independently, with the same parameters of our pipeline, and we identified about 100,000 SNVs for the first dataset and 11,500 SNVs for the second one (Figs. 3A–B). Despite the highly different coverage between the validation datasets and our dataset, we observed that the majority of called SNVs (95% for the first validation dataset and 93% for the second) were in common with the previously identified from our sequencing data (Fig. 3C). This large overlap excludes technical issues and sequencing errors. We also compared only the novel identified SNVs in our study with the novel identified SNV by using the validation datasets. Despite a slight reduction of the overlap percentage, indicating that the novel SNVs have a higher occurrence during long term cultures with respect to the known SNVs, the number of the common novel SNVs still remains high (~86% and ~81% respectively for Helin and Chen datasets) suggesting that the majority of the identified SNVs could be due to a different genetic background (Fig. S1A). To test this hypothesis, we also compared our identified SNVs with the SNVs obtained from the 129/Ola genome (http://www.sanger.ac.uk/resources/mouse/genomes/) analyzed with our pipeline. We found that more than 75% of the identified SNVs in E14 are present also in the 129/Ola strain mouse (Fig. 3D) demonstrating that most of the observed differences are due to a different genetic background between the mouse strain used for mm9 assembly (C57BL/6J) and the mouse strain used for derivation of E14 ESC (129/Ola) rather than differences acquired during different culture methods and/or times.

3.3. SNV functional classification

To classify the about 2.7×10^6 SNVs and to understand their biological relevance in the cellular homeostasis, we classified them using Annovar software, which provides functional annotations of the genetic variants [28]. We identified several thousands of potential functional SNVs residing in exons, splicing sites and in untranslated regions (UTR) of mRNA and in ncRNA. We found 8385 non-synonymous exonic SNVs that lead to an alteration in the amino acid sequence of the related protein. Moreover, we found 28 SNVs inserting an amiss stop codon and 18 SNVs leading the loss of the proper stop codon (Fig. 4A). Interestingly 30 of this loss or gain stop SNVs are present also in the 129/Ola strain. We found only 3 novel SNVs in mESC E14 expressed genes (Fig. S1B). The majority of the non-synonymous SNVs (about 70%) reside in no or low expressed gene and gene ontology analysis showed an enrichment in genes involved in biological processes as cell metabolism and cell adhesion (Figs. 4B–C). Development and stemness pathways do not seem to be affected suggesting the maintenance of the proper biological properties of this cell line.

We observed that the CpG dinucleotide shows higher variation frequency than other dinucleotides, probably due to the presence of methylation on 5′ of the cytosine in CpG context (Fig. 5A) [29,30]. For this reasons we compared the variation frequency in CpGs not residing in CpG islands that are more methylated in ESC (Fig. 5B) and we observed higher variation frequency in these CpGs with respect to those in CpG island context (Fig. 5C).

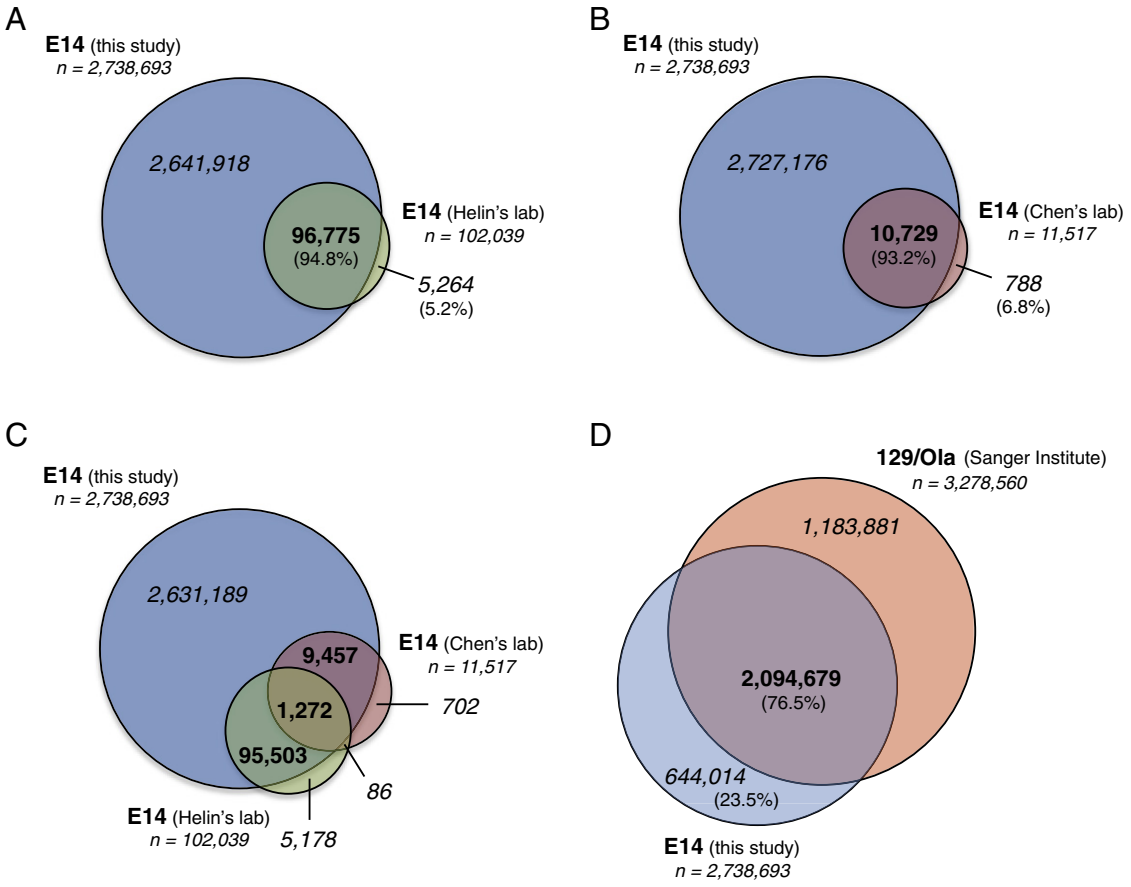


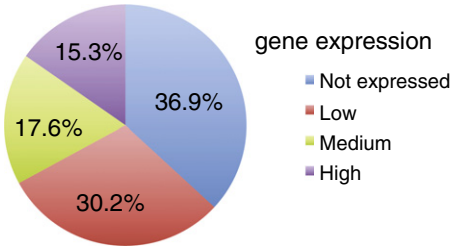
Fig. 3. A–B–C) Venn diagrams showing the overlap between the SNV dataset identified in this study and the other two identified using data derived from other studies in E14. D) Venn diagram showing the overlap between the SNV dataset identified in this study and the others identified using data derived from 129/Ola strain.

A

rank	region	n° SNV	percentage
1	exonic	27923	1.003
2	splicing	73	0.003
3	ncRNA	49529	1.779
4	UTR5	4539	0.163
5	UTR3	23506	0.844
6	intronic	1103838	39.641
7	upstream (1kb)	22748	0.817
8	downstream (1kb)	23042	0.827
9	intergenic	1529380	54.923

region	n° SNV	percentage
synonymous	19492	69.81
nonsynonymous	8385	30.03
stopgain	28	0.10
stoploss	18	0.06

B



C

nonsynonymous variants

Biological Process	P value
cell adhesion	3.53E-22
cellular process	2.55E-17
cell-cell adhesion	4.12E-17
cell communication	1.78E-13
signal transduction	1.46E-12
primary metabolic process	2.67E-10
metabolic process	5.34E-09
immune system process	1.04E-08
cell motion	6.86E-07

Fig. 4. A) Functional classification of the SNVs identified in this study using Annovar software. B) Venn diagram showing the fraction of non-synonymous variants for each gene group divided by relative expression in embryonic stem cell E14. C) Gene ontology analysis of the identified non-synonymous variants was performed using Panther software (http://www.pantherdb.org/) [33].

3.4. New E14 genome reference improves next generation data mapping

Finally, we tried to map other genome-wide experiments in E14 ESC, in particular a WCE-Seq, ChIP-Seq and MeDIP-Seq produced by other laboratories (SRA database: SRR867642, SRR867641 and SRR070936) using our new assembled genome as reference and we observed a higher mapping efficiency (compared to the mm9 assembly) with an increase of about 3–5% of mapped reads (Fig. 5D).

3.5. New E14 genome reference improves the accuracy of bisulfite sequencing experiments

The presence of higher variation frequency on CpG dinucleotide prompted us to focus our attention on methylation data derived from Bisulfite or Reduced Representation Bisulfite Sequencing (BS-Seq and RRBS-Seq). Bisulfite treatment converts unmethylated cytosine into uracil that is then transformed into thymidine during PCR amplification. Thus, the presence of C2T mutation alters the actual methylation status of the relative cytosines. For this reason, we performed an RRBS analysis [31] in E14 ESCs and performed the mapping and the analysis both on the reference mm9 genome and on the E14 genome (Fig. 5E) using BSMAP v2.74 software [32]. We observed an increase in the number of perfect reads mapped. More than 1×10^5 additional CpGs were called when using E14 genome with respect to the mm9 genome (Figs. 5F–G). Interestingly, beyond the increase of CpG calls, we were able to discard about 22,000 false CpG calls (Fig. 5H).

Thus our genome fits best the parameters for assembly of next generation data in E14 ESCs and can be useful for all scientists working in this cell line.

3.6. Utility of this study

We analyzed the next generation sequencing data generated in our laboratory to produce a database of SNVs in the genome of the widely used ESC line E14. Since the large use of this cell to understand the biology of the stemness and of the early development in vitro, we believe that our study will be very useful for all scientists related to this field. We built a database of about 2.7×10^6 high confidence SNVs that can be used to check eventual functional mutation or to create a new reference genome for mapping of genome-wide data. We found more than 8000 non-synonymous variants that lead to an alteration in the protein sequence providing a useful tool for all people that perform proteomic approach in this cell line. We detected several thousands of SNVs in ncRNA and untranslated regions (UTR) of the mRNA that could be responsible for the functional interaction between non-coding RNA and other RNA/DNA/proteins such as the recently discovered interactions between long intergenic non-coding RNA (LincRNA) and protein or the mRNA targeting by miRNA. A large number of SNVs were also found in regions outside the exons. These SNVs could be very important because they could reside in possible regulatory regions such as promoters, enhancers and splicing sites. Our database will be also important for all the scientists working on transcriptional regulation and mRNA processing in this line of embryonic stem cell. We demonstrated that the analysis genome-wide data (such as ChIP-Seq or MeDIP-Seq datasets) generated by both our or other laboratories could be improved by using our genome as reference for read mapping, since bioinformatical analysis optimization is a crucial step of these applications. A particular attention should be used in the case of genome-wide analysis finalized to the methylation studies that used the bisulfite treatment. Indeed bisulfite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected thus allowing to follow the sequencing step, the detection of the original methylation state. This kind of analysis can lead to wrong interpretations in case of SNVs residing in CpGs.

4. Discussion

A total of 30 Illumina runs were used to generate a sequence data to perform a genotyping of the E14 genome by sequencing. We used only data produced in our laboratory from the same cell line e from the same

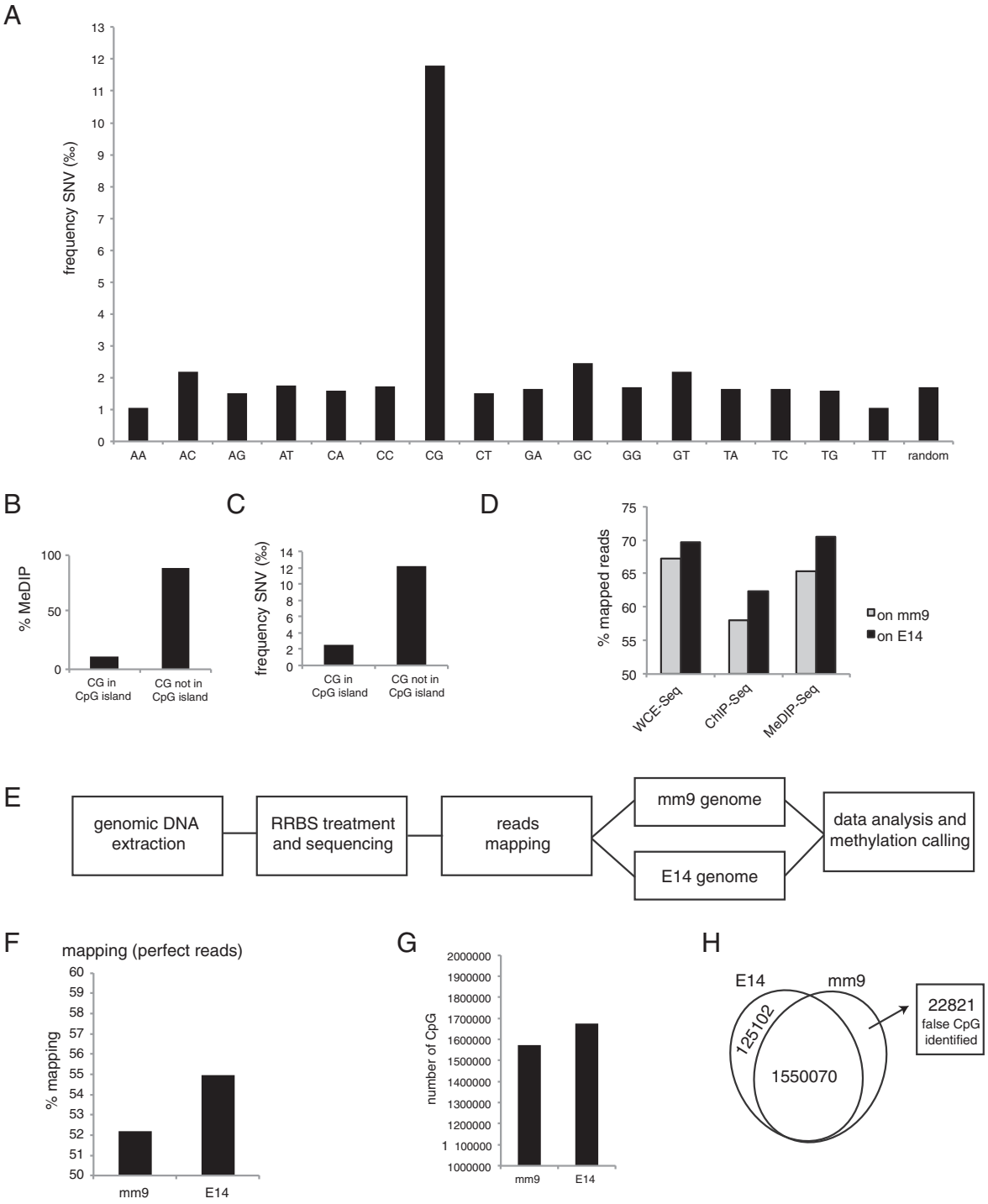


Fig. 5. A) SNV frequency (in ‰) around the genome of each dinucleotide. B) The majority of MeDIP peaks (methylated CpGs) reside in not CpG island context. C) methylated CpGs show higher variation frequency. D) Mapping with Bowtie on E14 genome assembly improves the number of reads correctly mapped. E) Overview of the RRBS experiment. F) Increased mapping efficiency of RRBS raw data on E14 genome with respect to the mm9 assembly. G–H) Number of CpGs called in RRBS experiment using the two different genome assembly mm9 or E14.

platform, to avoid to incorporate in the database mutations accumulated by culturing the cells for a long time. Nevertheless, we used next generation sequencing data from two other different laboratories and applied our pipeline to these data to discover the SNVs in their E14 cells. Because of the lower amount of sequence data available (which implies a lower coverage of the genome), the dataset of the identified SNV is smaller, but shows a large overlap with the datasets of the SNVs identified in our E14 cells. This overlap suggests that the relevant difference between the genotype of E14 and the mm9 genome deposited on NCBI or UCSC is due to different genetic background of the original mouse strains (129/Ola vs C57BL/6J), while the minimal difference (<5%) between the our and the other E14 cells could be due to different culture methods, times or aging of the cells. We found that the differences between different cultured E14 cells are enriched for common novel SNVs with respect to the annotated mouse SNVs. We built a single nucleotide variant database of E14 mouse embryonic stem cell. This database can be used for the study of individual functional mutations within genomic regulatory regions or within protein-coding

gene sequences. We produced a new E14 genome assembly that improves the mapping efficiency of next generation sequencing data that are made in this cell line. Given the high and increasing interest in mouse embryonic stem cell biology, and the increasing number of genome-wide studies, the genome assembly here produced should be a valuable tool for any researcher performing whole-genome analysis in embryonic stem cells. The 115 Gb of submitted E14 sequence datasets and the provided SNV database are a resource that can be used by all the scientific community working in E14 embryonic stem cells, both for genome-wide analysis and both for gene/transcript/protein specific analysis since many of the identified SNVs reside in functional (or potential) regions.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.06.007.

Additional material

Supplementary Table S1: list of the functional SNVs in the E14 genome.

The table includes the list of SNVs residing in exons of the coding/non-coding RNA, in splicing sites, in untranslated regions, upstream or downstream genes, and the functional classification of the identified SNV.

Database linking

The sequencing raw data are deposited as Fastq at GSE53149.

The list of the identified SNVs in E14 and the new assembled E14 genome can be downloaded from here: http://epigenetics.hugef-research.org/data.php.

The list of potentially functional SNVs in exons or regulatory regions is in Supplementary Table S1.

Competing interests

The authors declare that they have no competing, financial or non-financial interests.

References

[1] R.L. Gardner, F.A. Brook, Reflections on the biology of embryonic stem (ES) cells, Int. J. Dev. Biol. 41 (1997) 235–243.
[2] S.H. Orkin, K. Hochedlinger, Chromatin connections to pluripotency and cellular reprogramming, Cell 145 (2011) 835–850.
[3] C.E. Murry, G. Keller, Differentiation of embryonic stem cells to clinically relevant populations: lessons from embryonic development, Cell 132 (2008) 661–680.
[4] M. Hooper, K. Hardy, A. Handyside, S. Hunter, M. Monk, HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells, Nature 326 (1987) 292–295.
[5] A.G. Smith, M.L. Hooper, Buffalo rat liver cells produce a diffusible activity which inhibits the differentiation of murine embryonal carcinoma and embryonic stem cells, Dev. Biol. 121 (1987) 1–9.
[6] T. Seifert, S. Stoelting, T. Wagner, S.O. Peters, Vasculogeneic maturation of E14 embryonic stem cells with evidence of early vascular endothelial growth factor independency, Differentiation 76 (2008) 857–867.
[7] I. Ben-Porath, M.W. Thomson, V.J. Carey, R. Ge, G.W. Bell, A. Regev, et al., An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors, Nat. Genet. 40 (2008) 499–507.
[8] V. Clement, P. Sanchez, N. de Tribolet, I. Radovanovic, A. Ruiz i Altaba, HEDGEHOG-GLI1 signaling regulates human glioma growth, cancer stem cell self-renewal, and tumorigenicity, Curr. Biol. 17 (2007) 165–172.
[9] S.H. Chiou, M.L. Wang, Y.T. Chou, C.J. Chen, C.F. Hong, W.J. Hsieh, et al., Coexpression of Oct4 and Nanog enhances malignancy in lung adenocarcinoma by inducing cancer stem cell-like properties and epithelial–mesenchymal transdifferentiation, Cancer Res. 70 (2010) 10433–10444.
[10] C.Y. Darini, P. Martin, S. Azoulay, M.-D. Drici, P. Hofman, S. Obba, et al., Targeting cancer stem cells expressing an embryonic signature with anti-proteases to decrease their tumor potential, 42013. e706–e710.
[11] F. Neri, A. Zippo, A. Krepelova, A. Cherubini, M. Rocchigiani, S. Oliviero, Myc regulates the transcription of the PRC2 gene to control the expression of developmental genes in embryonic stem cells, Mol. Cell. Biol. 32 (2012) 840–851.
[12] S. Evellin, F. Galvagni, A. Zippo, F. neri, M. Orlandini, D. Incarnato, et al., FOSL1 controls the assembly of endothelial cells into capillary tubes by direct repression of αv and β3 integrin transcription, Mol. Cell. Biol. 33 (2013) 1198–1209.
[13] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells, Cell 133 (2008) 1106–1117.
[14] W.A. Pastor, U.J. Pape, Y. Huang, H.R. Henderson, R. Lister, M. Ko, et al., Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells, Nature 473 (2011) 394–397.
[15] K. Williams, J. Christensen, M.T. Pedersen, J.V. Johansen, P.A.C. Cloos, J. Rappsilber, et al., TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity, Nature 473 (2011) 343–348.
[16] F. Neri, A. Krepelova, D. Incarnato, M. Maldotti, C. Parlato, F. Galvagni, et al., Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs, Cell 155 (2013) 121–134.
[17] F. Neri, D. Incarnato, A. Krepelova, S. Rapelli, A. Pagnani, R. Zecchina, et al., Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells, Genome Biol. 14 (2013) R91.
[18] H. Marks, T. Kalkan, R. Menafra, S. Denissov, K. Jones, H. Hofemeister, et al., The transcriptional and epigenomic foundations of ground state pluripotency, Cell 149 (2012) 590–604.
[19] T.S. Mikkelsen, M. Ku, D.B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (2007) 553–560.
[20] A. Meissner, T.S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, et al., Genome-scale DNA methylation maps of pluripotent and differentiated cells, Nature 454 (2008) 766–770.
[21] M. Ku, R.P. Koche, E. Rheinbay, E.M. Mendenhall, M. Endoh, T.S. Mikkelsen, et al., Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains, PLoS Genet. 4 (2008) e1000242.
[22] A. Krepelova, F. neri, M. Maldotti, S. Rapelli, S. Oliviero, Myc and max genome-wide binding sites analysis links the myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells, PLoS ONE 9 (2014) e88933.
[23] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[24] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, et al., The Sequence Alignment/Map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[25] M.P. Creyghton, A.W. Cheng, G.G. Welstead, T. Kooistra, B.W. Carey, E.J. Steine, et al., Histone H3K27ac separates active from poised enhancers and predicts developmental state, Proc. Natl. Acad. Sci. 107 (2010) 21931–21936.
[26] N.D. Heintzman, G.C. Hon, R.D. Hawkins, P. Kheradpour, A. Stark, L.F. Harp, et al., Histone modifications at human enhancers reflect global cell-type-specific gene expression, Nature 459 (2009) 108–112.
[27] A. Visel, M.J. Blow, Z. Li, T. Zhang, J.A. Akiyama, A. Holt, et al., ChIP-seq accurately predicts tissue-specific activity of enhancers, Nature 457 (2009) 854–858.
[28] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (2010) e164.
[29] J. Xia, L. Han, Z. Zhao, Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome, BMC Genomics 13 (2012) S7.
[30] A. Hodgkinson, A. Eyre-Walker, Variation in the mutation rate across mammalian genomes, Nat. Rev. Genet. 12 (2011) 756–766.
[31] H. Gu, Z.D. Smith, C. Bock, P. Boyle, A. Gnirke, A. Meissner, Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling, Nat. Protoc. 6 (2011) 468–481.
[32] Y. Xi, W. Li, BSMAP: whole genome bisulfite sequence MAPping program, BMC Biol. 10 (2009) 232.
[33] H. Mi, A. Muruganujan, P.D. Thomas, PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees, Nucleic Acids Res. 41 (2013) D377–D386.