G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(1):34-40
http://dx.doi.org/10.5808/GI.2016.14.1.34

ORIGINAL ARTICLE

Non-negligible Occurrence of Errors in Gender Description in Public Data Sets

Jong Hwan Kim[1,2], Jong-Luyl Park[3], Seon-Young Kim[1,2]*

[1]Genome Structure Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Korea,
[2]Department of Functional Genomics, University of Science and Technology (UST), Daejeon 34113, Korea,
[3]Epigenome Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Korea

Due to advances in omics technologies, numerous genome-wide studies on human samples have been published, and most of the omics data with the associated clinical information are available in public repositories, such as Gene Expression Omnibus and ArrayExpress. While analyzing several public datasets, we observed that errors in gender information occur quite often in public datasets. When we analyzed the gender description and the methylation patterns of gender-specific probes (glucose-6-phosphate dehydrogenase [*G6PD*], ephrin-B1 [*EFNB1*], and testis specific protein, Y-linked 2 [*TSPY2*]) in 5,611 samples produced using Infinium 450K HumanMethylation arrays, we found that 19 samples from 7 datasets were erroneously described. We also analyzed 1,819 samples produced using the Affymetrix U133Plus2 array using several gender-specific genes (X (inactive)-specific transcript [*XIST*], eukaryotic translation initiation factor 1A, Y-linked [*EIF1AY*], and DEAD [Asp-Glu-Ala-Asp] box polypeptide 3, Y-linked [*DDDX3Y*]) and found that 40 samples from 3 datasets were erroneously described. We suggest that the users of public datasets should not expect that the data are error-free and, whenever possible, that they should check the consistency of the data.

Keywords: blood, DNA methylation, gender identity, gene expression, microarray analysis

Introduction

The completion of the human genome project has accelerated the development of many omics technologies that have been used extensively for the genomewide profiling of human samples [1]. Public repositories, such as Gene Expression Omnibus (GEO) and ArrayExpress, now hold data on hundreds of thousands of samples profiled by diverse technologies [2, 3]. For many of the samples in the public repositories, various kinds of clinical information are also included so that interested researchers can use them for their own research interests.

While these public data have enormous potential for research use, it is inevitable that unknown errors may creep into public datasets without being noticed by depositors. For example, Microsoft Excel is notorious for its automatic conversion function, which erroneously changes tens of human gene symbols [4]. But, the errors are not limited to gene symbols and may occur in the clinical information, as well.

Age and gender are some of the most basic clinical information associated with human samples and are most unlikely to be erroneous during data acquisition. While they are basic clinical information, the importance of age and gender in human biology is not trivial. There is strong evidence that men and women differ in terms of development and severity of many common diseases, including cardiovascular diseases, autoimmune diseases, and asthma [5]. Recent clinical studies have revealed an association between several genetic diseases and gender-specific genetic patterns [6, 7].

While analyzing public datasets produced using Infinium 450K HumanMethylation arrays (Illumina Inc., San Diego, CA, USA) and Affymetrix Human Genome U133Plus 2.0 arrays (Affymetrix Inc., Santa Clara, CA, USA), we found many samples that were discordant between clinical gender information and the patterns of gender-specific markers. Importantly, the errors were not limited to a few datasets but

Received September 9, 2015; Revised December 30, 2015; Accepted December 30, 2015

***Corresponding author:** Tel: +82-42-879-8116, Fax: +82-42-879-8119, E-mail: kimsy@kribb.re.kr

were prevalent in many datasets produced by many different laboratories. We advise that the users of public datasets should not expect that these data are error-free and, whenever possible, that they check the consistency of data.

Methods

Data collection and processing

Infinium 450K HumanMethylation array

More than 5,600 samples from 11 datasets were collected from GEO [2]. Before data integration, we calculated the average beta value for measuring methylation levels at each CpG site, ranging from 0 (least methylated) to 1 (most methylated). The individual beta value was dropped if the detection p-value was over 0.05. For all datasets, raw signal intensity files were collected, and from them, methylation level was calculated as $\beta = (\max(Cy5, 0)) / (|Cy3| + |Cy5| + 100)$. A constant of 100 was added to the denominator to regularize β values when both unmethylated and methylated intensities were small.

Affymetrix U133Plus2 array

A total of 1,819 samples from 4 datasets for gene expression experiments using the Affymetrix U133Plus2 array were obtained from GEO [2]. For all datasets, CEL files were collected and normalized by Robust Multiarray Average method [8].

Selection of gender-specific DNA methylation and gene expression markers

The gender-specific DNA methylation markers of the X chromosome were selected from reported X-linked housekeeping genes [9]. The gender-specific DNA methylation markers of the Y chromosome were estimated using differentially methylated CpG sites in Y chromosomes between males and females (Supplementary Fig. 1). Gender-specific markers were selected based on their beta value distribution in both males and females. As a result, two markers (cg24139739 and cg02869694) were selected as X chromosome markers, and two markers (cg07851521 and cg10835413) were selected as Y chromosome markers. The cutoff values for each marker were cg24139739 $<$ 0.25, cg02869694 $<$ 0.4, cg07851521 $>$ 0.5, and cg10835413 $>$ 0.45 (Supplementary Fig. 1).

The gender-specific gene expression markers were selected from reported gender-specific gene expression patterns in human blood (Supplementary Fig. 2) in a similar way [10]. Two markers (214218_s_at and 224588_at) were selected as X chromosome markers, and two markers (204409_s_at and 205000_at) were selected as Y chromosome markers. The cutoff values for each marker were 214218_s_at $<$ 4.5, 224588_at $<$ 7.5, 204409_s_at $>$ 5, and 205000_at $>$ 5.

Data analysis

Python version 2.7.6 and the Pandas python library version 0.15.2 were used for most data analyses. R version 3.1.0 and ggplot2 version 1.0.0 were used for image production.

Results

Collection of DNA methylation array data of human whole blood with age and gender information

We obtained DNA methylation microarray data from the NCBI GEO database (Fig. 1). We collected datasets of normal human blood samples in which both age and gender data were available. As a result, we collected a total of 4,862 samples for Infinium 450K HumanMethylation array data (Table 1).

Determination of gender by gender-specific markers

The CpG sites of X chromosomes in females are hypermethylated for dosage compensation of female X chromosomes [11]. By analyzing female-specific hypermethylated genes in the X chromosome (Supplementary

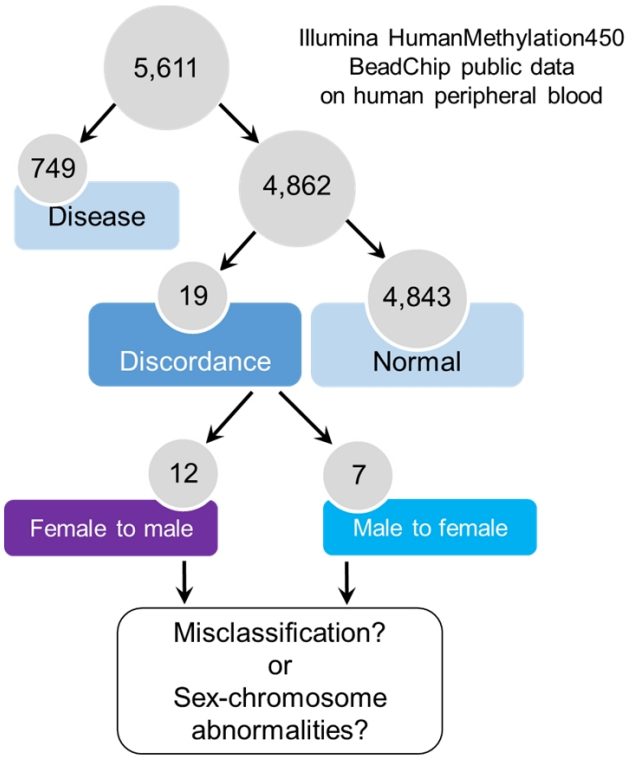

Fig. 1. Work flow of the Infinium 450K HumanMethylation array data analysis.

Table 1. Collected Infinium 450K HumanMethylation array samples and information on discordant samples

GSE	Raw data	Disease sample	Normal sample	Discordance sample	Age (y)	Sex	Predicted age	Residue
GSE32148	48	29	19	GSM796699	76	F	49.40	+26.59
GSE36064	78		78					
GSE40005	24	12	12					
GSE40279	656		656	GSM990316	59	M	62.60	−3.61
				GSM990443	84	F	73.93	+10.06
GSE41169	95	62	33	GSM1009739	29	F	26.61	+2.39
				GSM1009749	26	F	27.45	−1.45
GSE51032	845	421	424	GSM1235942	57.94	M	54.90	+3.04
				GSM1235964	54.31	F	58.91	−4.59
				GSM1236061	63.01	F	59.67	+3.34
				GSM1236313	58.54	F	38.89	+19.65
GSE53128	43		43	GSM1282801	66.78	F	50.21	+16.57
GSE53740	384	219	165	GSM1299660	68	M	71.01	−3.01
				GSM1299719	53	M	55.36	−2.36
				GSM1300551	86	F	80.42	+5.58
GSE55763	2,711		2,711	GSM1343079	72.9	F	66.84	+6.05
				GSM1343082	62.7	M	54.74	+7.95
				GSM1343115	72.9	F	71.30	+1.60
				GSM1343118	62.7	M	51.72	+10.98
				GSM1344329	60.4	M	73.65	−13.25
GSE56105	614		614					
GSE64495	113	6	107	GSM1572595	5.5	F (T)		

GSE, Gene Expression Omnibus (GEO) series; GSM, GEO sample; F, female; M, male; T, Turner syndrome.

Table 1), we identified 19 samples (0.39%) in which the methylation patterns of gender-specific CpG markers did not match the given gender information (Fig. 2A). When we analyzed those samples with male-specific hypermethylated genes in the Y chromosome (Supplementary Table 2), we again observed the opposite methylation patterns (Fig. 2B). Importantly, the discordant patterns were observed in eight of 11 datasets from different depositors, suggesting that the errors were not limited to one laboratory (Supplementary Fig. 3).

Interestingly, we found both types of discordant errors between DNA methylation patterns and given gender information. That is, for some samples, DNA methylation patterns were found to be female-specific while they were designated as males (designated here as discordant-male), and for the other samples, DNA methylation patterns were male-specific while they were designated as female (designated here as discordant-female).

Analysis of markers of chromosome abnormality syndromes

We next analyzed whether the observed discrepancy between methylation patterns of gender-specific genes and the given clinical information could be explained by rare sex chromosome abnormality syndromes, such as Turner and Klinefelter syndromes.

While a normal male inherits an X chromosome and a Y chromosome and a normal female inherits two X chromosomes, several abnormalities in the number of sex chromosomes occur gender-specifically. For females, abnormalities are a result of variations in the number of X chromosomes. For males, abnormalities are due to irregular numbers of the X or Y chromosome or both. The most frequent sex chromosome abnormalities in females are Turner (one X; 1:2,000) and triple X (XXX; 1:1,000), while those in males are Klinefelter (XXY; 1:500) and XYY (1:1,000) syndromes [12-15].

For comparison, we collected Infinium 450K HumanMethylation array data of one Turner syndrome patient and five Klinefelter syndrome patients. The characteristic methylation pattern of Turner syndrome (one X) is the hypomethylation of both X- and Y-specific markers (Fig. 3A). However, we found that most discordant-female samples in our dataset showed hypermethylation patterns in Y-specific markers, suggesting that they had XY chromosomes (a normal male) but not X (Turner syndrome) chromosome (Fig. 3A). Only one discordant-female sample (green circle in Fig. 2A) showed a pattern similar to Turner syndrome. For Klinefelter syndrome (XXY), the expected methylation patterns are the hypermethylation of both male- and

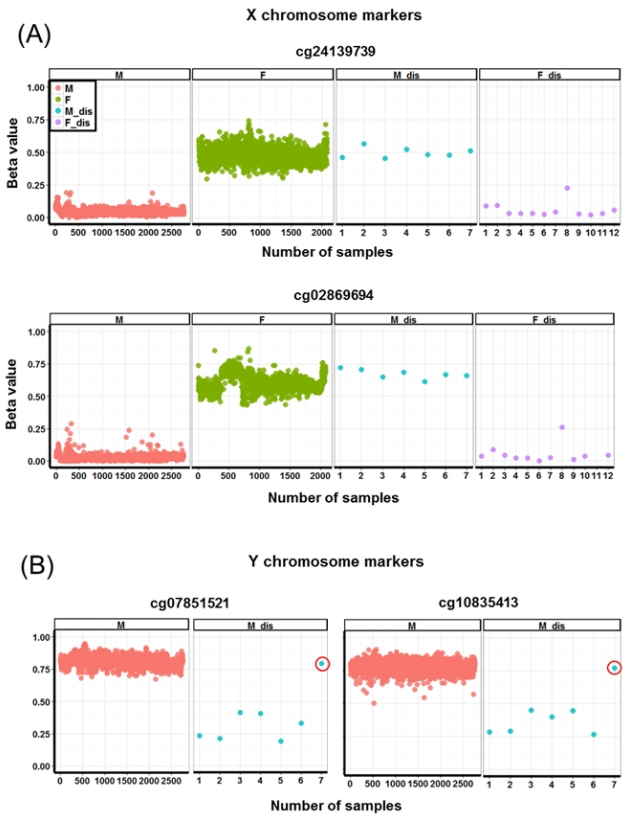

Fig. 2. Methylation pattern of gender-specific markers. X (A) and Y (B) chromosome markers. The red circles indicate the discordant male sample showing overlapping distribution of levels for normal male samples. M, male; F, female; M_dis, male samples showing methylation patterns of females; F_dis, female samples showing methylation patterns of males.

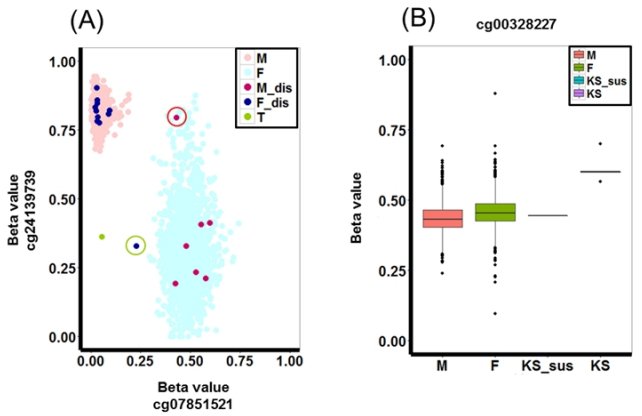

Fig. 3. Comparison of methylation status between normal and sex chromosome abnormalities. (A) Scatterplot of methylation status of gender-specific methylation markers on X and Y chromosomes. The red circle indicates the sample with suspected Klinefelter syndrome. The green circle indicates the sample with suspected Turner syndrome. (B) Box plot of Klinefelter syndrome-specific methylation markers of normal males, normal females, a sample with suspected Klinefelter syndrome, and Klinefelter syndrome samples. M, male; F, female; M_dis, male samples showing methylation patterns of females; F_dis, female samples showing methylation patterns of males; T, Turner syndrome sample; KS, Klinefelter syndrome sample; KS_sus, Klinefelter syndrome-suspected sample.

female-specific markers. We observed one discordant-male sample in which both male- and female-specific markers were hypermethylated (red circles in Figs. 2B and 3A). However, a close examination of three Klinefelter syndrome-specific markers revealed that the one discordant-male sample did not show the patterns of a Klinefelter syndrome patient (Fig. 3B, Supplementary Fig. 4) [16]. In conclusion, only one sample showed a methylation pattern associated with sex chromosome abnormalities.

Significant difference between clinical age information and predicted age using a DNA methylation age calculator

Recently, several studies have shown that DNA methylation markers can be used to estimate the age of an individual to within 5 y [17]. Because the datasets we analyzed had age information, as well, we tested whether the given age information deviated much from the estimated age from age-specific DNA methylation markers. The age prediction was performed by using the DNA methylation age calculator, a web-based tool that provides a predicted age based on the methylation values of DNA methylation markers of Illumina's Infinium platform [18]. When we compared the absolute age deviation between concordant and discordant samples, we found much larger deviations in age prediction among discordant samples compared with concordant samples (Table 1, Supplementary Fig. 5). This result suggests that the discordant samples are highly likely to have errors both in age and gender information, a scenario that occurs when two samples are mislabeled as each other.

Identification of gender-discordant samples from the analysis of gender-specific gene expression

To see if the same problem occurs in other types of data (e.g., gene expression), we collected gene expression microarray data from the NCBI GEO database using a similar strategy for methylation microarray data (Fig. 4). By applying the same criteria as with the DNA methylation data, we collected a total of 1,683 samples from 4 datasets produced using the Affymetrix U133Plus2 array platform (Table 2).

We analyzed the 1,683 samples using several gender-specific genes (X (inactive)-specific transcript [XIST], eukaryotic translation initiation factor 1A, Y-linked [EIF1AY], and DEAD [Asp-Glu-Ala-Asp] box polypeptide 3, Y-linked [*DDDX3Y*]) [10] and found that 40 samples (2.3%) from 3 datasets were erroneously described (Fig. 5). In the case of Y chromosome markers, some discordant male

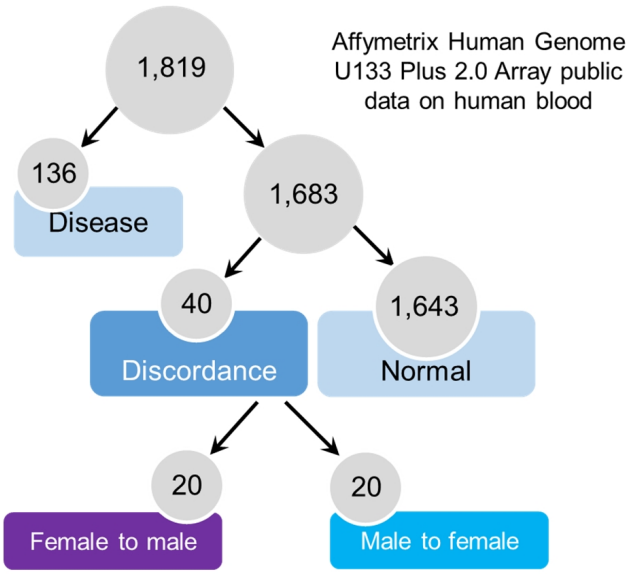

Fig. 4. Work flow of the Affymetrix Human Genome U133Plus 2.0 array analysis.

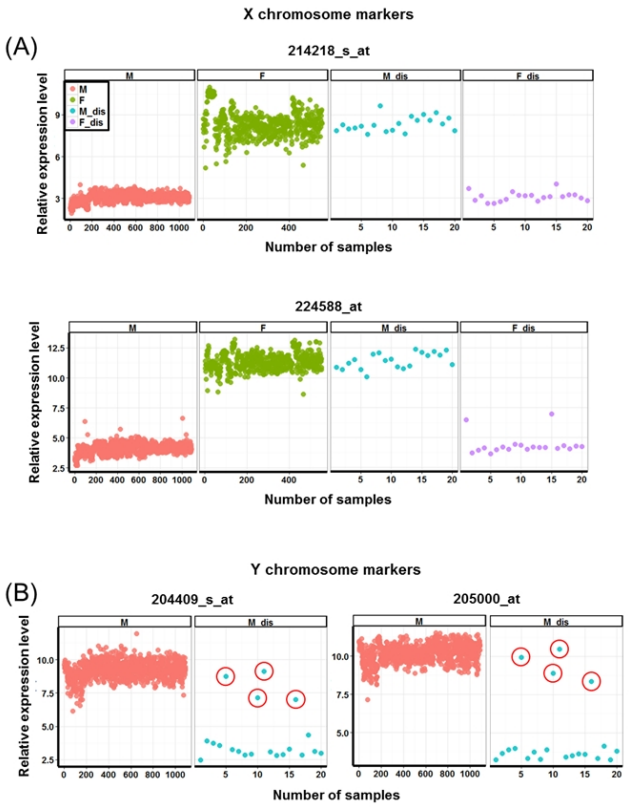

Fig. 5. Patterns of gender-specific gene expression of X- and Y-specific genes. X (A) and Y (B) chromosome markers. The red circles indicate the discordant male samples showing overlapping distribution of levels for normal male samples. M, male; F, female; M_dis, male samples showing methylation patterns of females; F_dis, female samples showing methylation patterns of males.

Table 2. Collected Affymetrix U133Plus2 array datasets and information on discordant samples

GSE	Raw data	Disease sample	Normal sample	Discordance sample	Sex
GSE13501	195	136	59		
GSE19743	177		177	GSM493658	M
				GSM493709	F
				GSM493727	M
				GSM493785	F
				GSM493820	F
				GSM493823	F
				GSM493824	F
				GSM493829	F
GSE36809	857		857	GSM901387	F
				GSM901391	F
				GSM901627	F
				GSM901682	M
				GSM901802	M
				GSM901827	F
				GSM902054	M
				GSM902061	M
GSE37069	590		590	GSM909660	F
				GSM909672	F
				GSM909697	M
				GSM909776	M
				GSM909777	M
				GSM909783	M
				GSM909802	M
				GSM909805	M
				GSM909811	M
				GSM909865	F
				GSM909894	F
				GSM909895	F
				GSM909907	F
				GSM909945	M
				GSM909949	M
				GSM909977	M
				GSM909983	M
				GSM910019	M
				GSM910103	F
				GSM910105	F
				GSM910151	M
				GSM910178	M
				GSM910184	F
				GSM910205	F

GSE, Gene Expression Omnibus (GEO) series; GSM, GEO sample; M, male; F, female.

samples showed overlapping distribution of methylation levels compared with normal male samples. Those samples needed to be compared samples with male sex abnormality syndromes (e.g., Klinefelter syndrome), but unfortunately, we could not find datasets with male sex abnormality syndromes produced using the same platform. However, as these samples showed the opposite pattern in X chromo-

some markers, we considered them to be putative discordant samples. But, further validation will be needed to confirm these samples.

Unlike the previous methylation data, some of the collected gene expression datasets contained technical replicates of the same samples (Table 2). Surprisingly, discordant gender-specific expression patterns were even observed from some of the technically replicated samples (Supplementary Fig. 6).

Discussion

Recently, genomewide omics data have accumulated very fast due to advances in omics technologies (e.g., Illumina's Infinium methylation assay and next-generation sequencing). Now, researchers can exploit public data repositories (e.g., ArrayExpress [3] and GEO [2]) that store data of hundreds of thousands of samples with both genomewide profiling data and associated clinical information. However, as we have shown here, datasets are not error-free, and any type of error can occur in public datasets. Indeed, one paper reported that the aggregate mislabeled blood sample rate was 1.12% at various US institutions in 2013 [19]. Possibly, various kinds of errors (e.g., mislabeling of tubes, mixing of samples) may occur from sample preparation to various steps of the experimental procedures.

When we began the analysis of Infinium 450K HumanMethylation array data, we did not expect errors in gender description to occur so often in several datasets. We thus analyzed if those erroneous samples showed the methylation patterns of gender-specific chromosome abnormality syndromes (e.g., Turner and Klinefelter syndromes) but found that most of the erroneous samples did not. Thus, we concluded that most discordant samples were real human errors. This finding led us to analyze expression datasets of Affymetrix U133Plus2 array data, as well, and again, we found that errors were not rare as well. Fortunately, for transcriptomic and epigenomic datasets, gender-specific markers (Supplementary Tables 1–4) are well known, so that researchers can check whether the given gender information is concordant with the expected patterns of gender-specific markers. For DNA methylation, a user-friendly website (https://dnamage.genetics.ucla.edu/) [18] is also available for researchers.

In conclusion, we suggest that users of public data should not expect that the data are error-free, and whenever possible, they should check them carefully before use.

Supplementary materials

Supplementary data including six figures and four tables can be found with this article online at http://www.genominfo.org/src/sm/gni-14-34-s001.pdf.

Acknowledgments

This work was supported by the Forensic Science Research Project 2014 of the Supreme Prosecutors' Office and a KRIBB research initiative grant.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, *et al*. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res* 2009;37:D885-D890.
3. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, *et al*. ArrayExpress: a public database of gene expression data at EBI. *C R Biol* 2003;326:1075-1078.
4. Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, *et al*. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics* 2004;5:80.
5. Ober C, Loisel DA, Gilad Y. Sex-specific genetic architecture of human disease. *Nat Rev Genet* 2008;9:911-922.
6. Van der Meulen J, Sanghvi V, Mavrakis K, Durinck K, Fang F, Matthijssens F, *et al*. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. *Blood* 2015;125:13-21.
7. Dimas AS, Nica AC, Montgomery SB, Stranger BE, Raj T, Buil A, *et al*. Sex-biased genetic effects on gene regulation in humans. *Genome Res* 2012;22:2368-2375.
8. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;4:249-264.
9. Bibikova M, Chudin E, Wu B, Zhou L, Garcia EW, Liu Y, *et al*. Human embryonic stem cells have a unique epigenetic signature. *Genome Res* 2006;16:1075-1083.
10. Guillén IA, Fernández JR, Palenzuela DO, Dueñas S, Han J, Zhang Z, *et al*. Analysis of gene expression profile for gender in human blood samples. *Int J Innov Appl Stud* 2014;7:329-342.
11. Mohandas T, Sparkes RS, Shapiro LJ. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* 1981;211:393-396.
12. Donaldson MD, Gault EJ, Tan KW, Dunger DB. Optimising management in Turner syndrome: from infancy to adult transfer. *Arch Dis Child* 2006;91:513-520.
13. Nielsen J, Wohlert M. Chromosome abnormalities found among 34,910 newborn children: results from a 13-year incidence study in Arhus, Denmark. *Hum Genet* 1991;87:81-83.
14. Otter M, Schrander-Stumpel CT, Curfs LM. Triple X syndrome: a review of the literature. *Eur J Hum Genet* 2010;18:265-271.

15. Stochholm K, Bojesen A, Jensen AS, Juul S, Gravholt CH. Criminality in men with Klinefelter's syndrome and XYY syndrome: a cohort study. *BMJ Open* 2012;2:e000650.
16. Wan ES, Qiu W, Morrow J, Beaty TH, Hetmanski J, Make BJ, *et al*. Genome-wide site-specific differential methylation in the blood of individuals with Klinefelter syndrome. *Mol Reprod Dev* 2015;82:377-386.
17. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, *et al*. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell* 2013;49:359-367.
18. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol* 2013;14:R115.
19. Grimm E, Friedberg RC, Wilkinson DS, AuBuchon JP, Souers RJ, Lehman CM. Blood bank safety practices: mislabeled samples and wrong blood in tube: a Q-Probes analysis of 122 clinical laboratories. *Arch Pathol Lab Med* 2010;134:1108-1115.