Genomics Data 9 (2016) 18–21

Contents lists available at ScienceDirect

Genomics Data

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

Integrity of genome-wide genotype data from low passage lymphoblastoid cell lines

Nina S. McCarthy [a,b,c], Spencer M. Allan [a], David Chandler [b,1], Assen Jablensky [b,c], Bharti Morar [b,c,d,*]

[a] Centre for the Genetic Origins of Health and Disease, The University of Western Australia, Perth, Australia
[b] Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, Australia
[c] Cooperative Research Centre for Mental Health, Carlton South, Victoria, Australia
[d] Harry Perkins Institute of Medical Research and Centre for Medical Research, The University of Western Australia, Perth, Australia

ARTICLE INFO

Article history:
Received 25 February 2016
Received in revised form 9 May 2016
Accepted 9 May 2016
Available online 12 May 2016

Keywords:
Lymphoblastoid cell line
Genotyping
Single nucleotide polymorphism

ABSTRACT

We compared genotype data from the HumanExomeCore Array in peripheral blood mononuclear cells and low passage lymphoblastoid cell lines from the same 24 individuals to test for genotypic errors caused by the Epstein–Barr Virus transformation process. Genotype concordance across the 24 comparisons was 99.57% for unfiltered genotype data, and 99.63% following standard genotype quality control filters. Mendelian error rates and levels of heterozygosity were not significantly different between lymphoblastoid cell lines and their parent peripheral blood mononuclear cells. These results show that at low passage numbers, genotype discrepancies are minimal even before stringent quality control, and extend current evidence qualifying the use of low-passage lymphoblastoid cell lines as a reliable DNA source for genotype analysis.

1. Introduction

Lymphoblastoid cell lines (LCLs), which are human B lymphocytes immortalized by in vitro infection with Epstein-Barr Virus (EBV), are a renewable source of DNA, and an alternative to primary cells or tissue samples as a source of genomic DNA. More and more DNA will be required as the genomics era progresses from genome-wide association studies to whole exome sequencing and whole genome sequencing, and studies are likely to utilise diverse sets of samples, possibly comprising combinations of DNA from peripheral blood mononuclear cells (PBMC) and LCLs.

A number of studies have shown that the immortalisation and/or subsequent passaging of these LCLs may lead to new mutations and genomic instability, including extended homozygosity, structural genomic variation and changes in DNA methylation patterns [1–8]. These new non-germline mutations would confound association studies of human disease, especially as the field moves towards rare variant analysis. Other studies have however indicated that LCLs with low passage numbers display good genomic stability, with the EBV-transformation process producing minor, if any, artefacts on genomic structure [9,10]. Two recent studies have reported high genotype concordance between DNA from LCLs and their parent PBMCs at low passage numbers (> 99%), especially after genotype quality control filtering has been applied [7,11], though there is evidence that high numbers of cell passages may produce instability [7]. Similarly, a small number of next-generation sequencing studies have reported that stringent filtering parameters significantly reduce discordant calls and validation experiments indicate minimal differences between PBMC–LCL pairs [6,10,12].

To add strength to the findings of the limited number of studies that have assessed the validity of using LCL DNA in genetic studies, we have tested for genotypic errors potentially induced by the EBV transformation process by comparing single nucleotide polymorphism (SNP) genotype calls in PBMCs and LCLs from the same individuals ($N = 24$). Our cohort included two family groups, allowing for the detection of Mendelian errors. All samples were at early passage (< 5) and were genotyped on the Illumina HumanExomeCore Array, which contains > 500,000 common and 'rarish' SNPs. We report high concordance between PBMC–LCL pairs, and contrary to previous studies, our data do not show marked overall improvement in concordance after application of genotype quality control filtering. These data support the use of low passage LCL DNA in genetic studies where PBMC DNA from an individual is unavailable/depleted.

2. Materials and methods

2.1. Sample collection and generation of LCLs

The study sample comprised 24 individuals from the Western Australian Family Study of Schizophrenia [13] (WAFSS), including 16

* Corresponding author at: Centre for Clinical Research in Neuropsychiatry, QEII Medical Centre, 6 Verdun Street, Nedlands, WA 6009, Australia.
E-mail address: bharti.morar@perkins.uwa.edu.au (B. Morar).
[1] Present address: Australian Genome Research Facility Ltd., Perth, Australia.

http://dx.doi.org/10.1016/j.gdata.2016.05.006

unrelated individuals and two nuclear families — one trio, and one family consisting of parents and 3 offspring, DNA was extracted using standard protocols and stored frozen in 1XTE buffer. LCLs were generated as described in Verbrugghe et al. [14] Briefly, lymphocytes were isolated from whole blood samples using Ficoll Lymphocyte Separation Medium (MP Biochemicals). LCLs were established by transformation of fresh lymphocytes with EBV and cultured in a T25 flask in advanced RPMI medium supplemented with 2% fetal calf serum, 2 mM L-glutamax, 50 units/ml penicillin/50 μg/ml streptomycin and 2% crude phytohemagglutinin (M Form) [Invitrogen, Carlsbad, CA, USA] in a humidified environment at 37 °C in 5% (v/v) carbon dioxide. The culture was maintained in this media until it reached a cell density of 0.5–1 × 10^6 cells/ml in 20 ml. The cells were then transferred to a T75 flask and allowed to reach a density of ~1 × 10^6 cells/ml in 50–55 ml. At this stage, an aliquot of cells was removed for DNA isolation and the remaining cells cryopreserved in duplicate in liquid nitrogen. DNA was isolated from the cells using standard protocols and stored at −80 °C.

The study was approved by the Human Research Ethics Committee of The University of Western Australia. Written informed consent was obtained from all subjects.

2.2. Genotyping

Genotyping on the Illumina HumanCoreExome beadchip-12v1-1_A was performed at Pathwest, QEII Medical Centre (Nedlands, WA, Australia) according to manufacturer's instructions. The chip assays approximately 250,000 common [minor allele frequency (MAF) > 5%] and 250,000 'rarish' exonic SNPs (MAF 1–5%); details on the history and content of the chip can be found at http://genome.sph.umich.edu/wiki/Exome_Chip_Design.

2.3. Quality control

For the 'unfiltered' analysis, all 542,585 SNPs on the chip were compared between the 24 PBMC–LCL pairs. In order to assess whether concordance levels were improved following standard genotyping quality control measures, the quality control filters described in Table 1 were applied to the genotype data using PLINK [15] (http://pngu.mgh.harvard.edu/purcell/plink/). Rates of SNP heterozygosity were also calculated for each sample across the 232,171 autosomal markers which were polymorphic in this population. In addition, to provide a baseline error rate for this assay, 6 PBMC samples were genotyped in duplicate on the same 542,585 SNPs.

2.4. Statistical analysis

Tests of equal proportions were performed using the two-Sample test for equality of proportions as implemented in the prop.test function in the R package 'stats', or the paired test pairwise.prop.test with correction for multiple testing when comparing groups.

2.5. Results

Call rate as a proportion of the 542,585 SNPs on the chip was significantly lower overall for PBMCs than for LCLs ($P < 2.2 \times 10^{-16}$; Table 1) and remained significant after adjusting for between-sample variability (paired test of proportions $P = 0.001$). Application of the control filters (Table 1) to PBMC and LCL datasets resulted in a significantly different proportions of SNPs being removed due to MAF filtering and missingness. Following these exclusions, 237,429 and 239,448 SNPs remained in the PBMC and LCL datasets, respectively. Call rate was significantly different between most individual PBMC–LCL pairs ($P < 0.05$; Table 2 and Fig. 1), whereas genome wide rates of heterozygosity and number of Mendelian errors based on the filtered data were not significantly different between individual pairs (Table 2), or overall.

Genotype concordance between individual PBMC–LCL pairs was high across unfiltered (range 0.969–1.000, mean = 0.996, SD = 0.007) and QC filtered (range 0.979–0.998, mean = 0.996, SD = 0.004) datasets (Table 2 and Fig. 1). However, concordance between each individual pair for unfiltered and filtered SNP sets was significantly different (Table 2), although the direction of effect varied between samples. On average, there was a non-significant increase in concordance across all 24 pairs following quality control filtering (paired *t*-test, $P = 0.715$). By comparison, genotyping rate was 99.21% in the 6 PBMCs genotyped in duplicate (12 samples in total). Concordance between the genotypes in each replicate pair was 100%.

There were no associations with sample age or sex for any of the quality control measures or concordance rates (linear/logistic regression, $P > 0.05$).

3. Discussion

This study provides further evidence for minimal rates of discordant genotypes between PBMC and LCL pairs at low passage numbers, supporting the use of low-passage LCLs as a reliable DNA source for genotype analysis. Contrary to previous reports, there was no significant increase in concordance rates after stringent quality control filtering of the genotype data. We were able to check Mendelian error rates in

Table 1
Sample (upper panel) and SNP (lower panel) quality control exclusions for the 24 PBMC-LCL pairs. *P* values are for 2-sample tests for equality of proportions between the PBMC and LCL values. MAF — minor allele frequency; HWE — Hardy Weinberg equilibrium.

Samples			
Total samples			24
Sample exclusions			
Genotypes inconsistent with phenotypic sex			0
Samples with >10% of SNP genotypes missing			0
Samples with >5% SNPs showing Mendelian errors			0
Final samples remaining after exclusions			24
SNPs	PBMCs ($n = 24$)	LCLs ($n = 24$)	*P*
Total SNPs	542,585	542,585	
Average call rate (sd; range) %	99.10 (1.08; 95.90–99.90)	99.90 (0.08; 99.50–99.90)	< 2.2e − 16
SNP exclusions			
SNPs with >10% of genotypes missing	11,073 (2.04%)	959 (0.18%)	< 2.2e − 16
SNPs remaining with ≥ 90% genotype rate	*531,512*	*541,626*	
SNPs with minor allele frequency (MAF) < 5%	294,074 (55.32%)	302,164 (55.79%)	1.2E − 06
SNPs remaining with MAF ≥ 5%	*237,438*	*239,462*	
SNPs not in HWE (<$P = 0.001$)	0	0	NA
SNPs remaining in HWE (≥$P = 0.001$)	*237,438*	*239,462*	
SNPs with >10% Mendel error rate	9 (0.004%)	14 (0.006%)	0.41
SNPs remaining with ≤ 90% Mendel error rate	*237,429*	*239,448*	
Final SNPs remaining after exclusions	237,429 (43.76%)	239,448 (44.13%)	9.5E − 05

Table 2
Comparative data for individual PBMC–LCL pairs analysed in the study ($N = 24$). The first panel shows the comparison in call rate in the unfiltered ($n = 542.585$) SNP set. The second panel shows a comparison of heterozygosity levels based on QC-filtered autosomal SNPs common to both PBMC and LCL datasets ($n = 232.171$). In the third panel, Mendelian errors are reported for the two nuclear families present in the sample — one trio (FID F_3), and one family consisting of parents and 3 offspring (FID F_15). For these first three panels, *P* values are for 2-sample test for equality of proportions between individual PBMC and LCL pairs. The fourth panel shows concordance rates between PBMC–LCL pairs before and after QC filtering of SNPs. *P* values are for 2-sample test for equality of proportions (concordance) between unfiltered and filtered data. IID: individual ID; FID: Family ID; PID: paternal ID; MID: maternal ID; Sex: M — male, F — female; Age: age of the individual at the time of blood collection.

						% call rate (nSNPs = 542,585)			% heterozygosity (nSNPs = 232,171)			Mendelian errors (nSNPs 237,429)			Concordance rate between PBMC–LCL pairs		
IID	FID	PID	MID	SEX	AGE	PBMC	LCL	*P*	PBMC	LCL	*P*	PBMC	LCL	*P*	Unfiltered, nSNPs = 542,585	QC filtered, nSNPs = 237 .429	*P*
1	F_1	25	26	M	31	0.983	0.999	<2e−16	0.379	0.380	0.834	—	—	—	0.995	0.996	<2.2e−16
2	F_2	27	28	M	33	,	0.999	<2e−16	0.381	0.381	0.946	—	—	—	0.995	0.996	9.8e−15
3	F_3	5	4	F	34	0.999	0.999	1.5E−03	0.428	0.428	0.678	58	60	1	1.000	0.998	<2.2e−16
4	F_3	—	—	F	62	0.999	0.999	8.5E−01	0.395	0.394	0.674	31	32	1	1.000	0.998	<2.2e−16
5	F_3	—	—	M	64	0.999	0.999	2.7E−05	0.456	0.455	0.630	29	30	1	1.000	0.998	<2.2e−16
6	F_4	29	30	M	25	0.983	0.999	<2e−16	0.376	0.376	0.889	—	—	—	0.995	0.997	<2.2e−16
7	F_5	31	32	M	26	0.999	0.999	6.9E−01	0.378	0.378	0.946	—	—	—	0.999	0.998	<2.2e−16
8	F_6	33	34	M	31	0.981	0.995	<2e−16	0.377	0.377	0.946	—	—	—	0.992	0.996	<2.2e−16
9	F_7	35	36	M	37	0.999	0.999	3.9E−05	0.379	0.379	0.946	—	—	—	1.000	0.998	<2.2e−16
10	F_8	37	38	M	28	0.999	0.999	1.2E−01	0.380	0.380	0.889	—	—	—	1.000	0.998	<2.2e−16
11	F_9	39	40	M	20	0.982	0.999	<2e−16	0.372	0.372	0.946	—	—	—	0.993	0.997	<2.2e−16
12	F_10	41	42	M	25	0.981	0.999	<2e−16	0.380	0.380	1.000	—	—	—	0.993	0.996	<2.2e−16
13	F_11	43	44	M	33	0.999	0.999	1.3E−01	0.380	0.380	0.889	—	—	—	1.000	0.998	<2.2e−16
14	F_12	45	46	M	30	0.959	0.999	<2e−16	0.380	0.381	0.437	—	—	—	0.969	0.979	<2.2e−16
15	F_13	47	48	F	45	0.999	0.999	6.7E−01	0.382	0.381	0.946	—	—	—	1.000	0.998	<2.2e−16
16	F_14	49	50	F	38	0.979	0.999	<2e−16	0.376	0.375	0.621	—	—	—	0.986	0.987	4.2e−05
17	F_15	19	18	M	23	0.983	0.999	<2e−16	0.416	0.416	0.782	61	64	1	0.994	0.997	<2.2e−16
18	F_15	—	—	F	55	0.983	0.999	<2e−16	0.416	0.416	0.947	61	60	1	0.995	0.997	<2.2e−16
19	F_15	—	—	M	54	0.983	0.999	<2e−16	0.416	0.416	0.891	57	52	1	0.995	0.997	<2.2e−16
20	F_15	19	18	M	27	0.999	0.999	5.0E−01	0.414	0.414	0.891	27	19	1	1.000	0.998	<2.2e−16
21	F_15	19	18	F	25	0.999	0.999	3.2E−01	0.415	0.415	0.891	26	26	1	1.000	0.998	<2.2e−16
22	F_16	51	52	F	32	0.999	0.999	2.1E−01	0.381	0.381	1.000	—	—	—	0.999	0.998	<2.2e−16
23	F_16	51	52	F	35	0.999	0.999	1.0E+00	0.384	0.384	0.946	—	—	—	1.000	0.998	<2.2e−16
24	F_17	53	54	M	20	0.999	0.999	1.9E−05	0.378	0.377	0.889	—	—	—	1.000	0.998	<2.2e−16

our two family groups, and report comparable rates of Mendelian error in PBMC and LCL DNA. Surprisingly, we report significantly higher genotype call rates in the LCL DNA, which may indicate some degradation of the PBMC DNA.

Conflict of interest

The authors declare no conflict of interest.

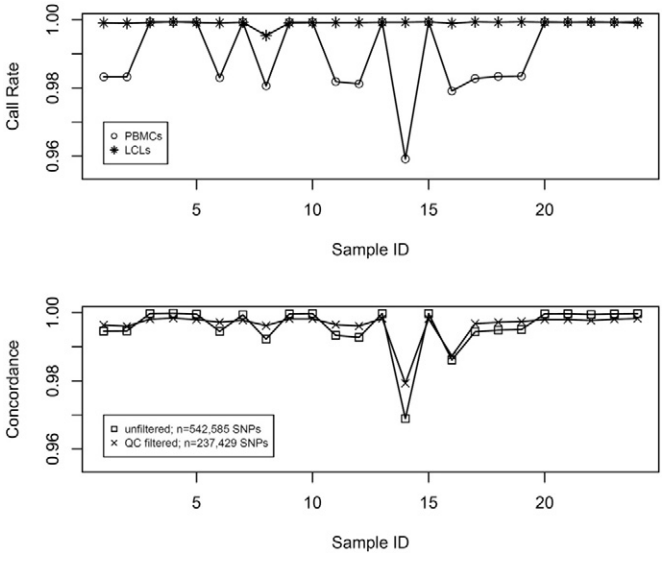

Fig. 1. Upper panel: Call rates for PBMCs and LCLs across all genotyped SNPs ($N = 542.585$). Lower panel: concordance rates between PBMC and LCL genotypes for all SNPs before and after QC filtering (237,429 SNPs common to both datasets after filtering).

Acknowledgements

We thank patients, family members and volunteer controls for their participation. The study was supported by Grants #37580400 and #1064582 from the National Health and Medical Research Council of Australia to Professor A. Jablensky, with funding contribution from the North Metropolitan Health Area, Perth, Western Australia and the Cooperative Research Centre (CRC) for Mental Health.

References

[1] R. Redon, S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, et al., Global variation in copy number in the human genome. Nature 444 (2006) 444–454.
[2] J. Simon-Sanchez, S. Scholz, H.C. Fung, M. Matarin, D. Hernandez, J.R. Gibbs, et al., Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum. Mol. Genet. 16 (2007) 1–14.
[3] D. Grafodatskaya, S. Choufani, J.C. Ferreira, D.T. Butcher, Y. Lou, C. Zhao, et al., EBV transformation and cell culturing destabilizes DNA methylation in human lymphoblastoid cell lines. Genomics 95 (2010) 73–83.
[4] S. Lacoste, E. Wiechec, A.G. Dos Santos Silva, A. Guffei, G. Williams, M. Lowbeer, et al., Chromosomal rearrangements after ex vivo Epstein–Barr virus (EBV) infection of human B cells. Oncogene 29 (2010) 503–515.
[5] D.F. Conrad, J.E. Keebler, M.A. DePristo, S.J. Lindsay, Y. Zhang, F. Casals, et al., Variation in genome-wide mutation rates within and between human families. Nat. Genet. 43 (2011) 712–714.
[6] E.R. Londin, M.A. Keller, M.R. D'Andrea, K. Delgrosso, A. Ertel, S. Surrey, et al., Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. BMC Genomics 12 (2011) 464.
[7] J.H. Oh, Y.J. Kim, S. Moon, H.Y. Nam, J.P. Jeon, J.H. Lee, et al., Genotype instability during long-term subculture of lymphoblastoid cell lines. J. Hum. Genet. 58 (2013) 16–20.
[8] M.D. Shirley, Concerns regarding "whole exome sequencing reveals minimal differences between cell line and whole blood derived DNA". Genomics 102 (2013) 430.
[9] J.P. Jeon, S.M. Shim, H.Y. Nam, S.Y. Baik, J.W. Kim, B.G. Han, Copy number increase of 1p36.33 and mitochondrial genome amplification in Epstein–Barr virus-transformed lymphoblastoid cell lines. Cancer Genet. Cytogenet. 173 (2007) 122–130.
[10] D. Nickles, L. Madireddy, S. Yang, P. Khankhanian, S. Lincoln, S.L. Hauser, et al., In depth comparison of an individual's DNA and its lymphoblastoid cell line using whole genome sequencing. BMC Genomics 13 (2012) 477.

[11] J.T. Herbeck, G.S. Gottlieb, K. Wong, R. Detels, J.P. Phair, C.R. Rinaldo, et al., Fidelity of SNP array genotyping using Epstein Barr virus-transformed B-lymphocyte cell lines: implications for genome-wide association studies. PLoS One 4 (2009), e6915.
[12] C.M. Schafer, N.G. Campbell, G. Cai, F. Yu, V. Makarov, S. Yoon, et al., Whole exome sequencing reveals minimal differences between cell line and whole blood derived DNA. Genomics 102 (2013) 270–277.
[13] A. Jablensky, Subtyping schizophrenia: implications for genetic research. Mol. Psychiatry 11 (2006) 815–836.
[14] P. Verbrugghe, S. Bouwer, S. Wiltshire, K. Carter, D. Chandler, M. Cooper, et al., Impact of the Reelin signaling cascade (Ligands-Receptors-Adaptor Complex) on cognition in schizophrenia. Am. J. Med. Genet. B Neuropsychiatr. Genet. 159B (2012) 392–404.
[15] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81 (2007) 559–575.