Human Molecular Genetics, 2009, Vol. 18, No. 20 ***3864–3875***
doi:10.1093/hmg/ddp330
Advance Access published on July 23, 2009

Expression differences by continent of origin point to the immortalization process

Adam R. Davis[1,*] and Isaac S. Kohane[1,2]

[1]i2b2 National Center for Biomedical Computing, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA and [2]Harvard Medical School Center for Biomedical Informatics, Boston, MA, USA

Received April 15, 2009; Revised June 15, 2009; Accepted July 16, 2009

Analysis of recently available microarray expression data sets obtained from immortalized cell lines of the individuals represented in the HapMap project have led to inconclusive comparisons across cohorts with different ancestral continent of origin (ACOO). To address this apparent inconsistency, we applied a novel approach to accentuate population-specific gene expression signatures for the CEU [homogeneous US residents with northern and western European ancestry (HapMap samples)] and YRI [homogenous Yoruba people of Ibadan, Nigeria (HapMap samples)] trios. In this report, we describe how four independent data sets point to the differential expression across ACOO of gene networks implicated in transforming the normal lymphoblast into immortalized lymphoblastoid cells. In particular, Werner syndrome helicase and related genes are differentially expressed between the YRI and CEU cohorts. We further demonstrate that these differences correlate with viral titer and that both the titer and expression differences are associated with ACOO. We use the 14 genes most differentially expressed to construct an ACOO-specific 'immortalization network' comprised of 40 genes, one of which show significant correlation with genomic variation (eQTL). The extent to which these measured group differences are due to differences in the immortalization procedures used for each group or reflect ACOO-specific biological differences remains to be determined. That the ACOO group differences in gene expression patterns may depend strongly on the process of transforming cells to establish immortalized lines should be considered in such comparisons.

INTRODUCTION

Several recent studies of populations of different ancestral continent of origin (ACOO) have identified ACOO-specific gene expression differences. Because the sets of genes identified in these studies are largely non-overlapping, the biological interpretation of these results is challenging (1–6). Given the importance to health disparities of such studies, we have undertaken an integrative approach to determine whether indeed there is a consistent difference. We have also added a new study sample to further validate our findings. Cross-population expression studies are fraught with the well-known variability in the biology as well as the difficulties in comparing transcriptome-wide measures from different platforms (7,8) and the increasingly documented intrinsic biases of expression patterns of immortalized cell lines (6). Technical bias may affect many genes in concert, thus causing spurious correlations in clinical data sets and false associations between genes and clinical variables (9). The study of the transcriptome in groups with different ACOO is particularly problematic in that most of these studies are performed on Epstein–Barr virus (EBV) immortalized cell lines. Specifically, the International HapMap Project harvested peripheral blood lymphoblasts from the homogenous Yoruba tribe from Ibadan Nigeria (YRI) and then transformed them into immortalized cells *in vitro* using the EBV. This is of potential additional relevance, as the YRI population is one of the sub-Saharan populations known to suffer from an endemic childhood cancer Burkitt lymphoma (BL), caused by the EBV that environmentally saturates sub-Saharan Africa (10–13). In contrast, the CEU [homogeneous US residents with northern and western European ancestry (HapMap samples)] population as well as

*To whom correspondence should be addressed at: i2b2 National Center for Biomedical Computing, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. Tel: +1 6173552933; E-mail: ardavis@partners.org

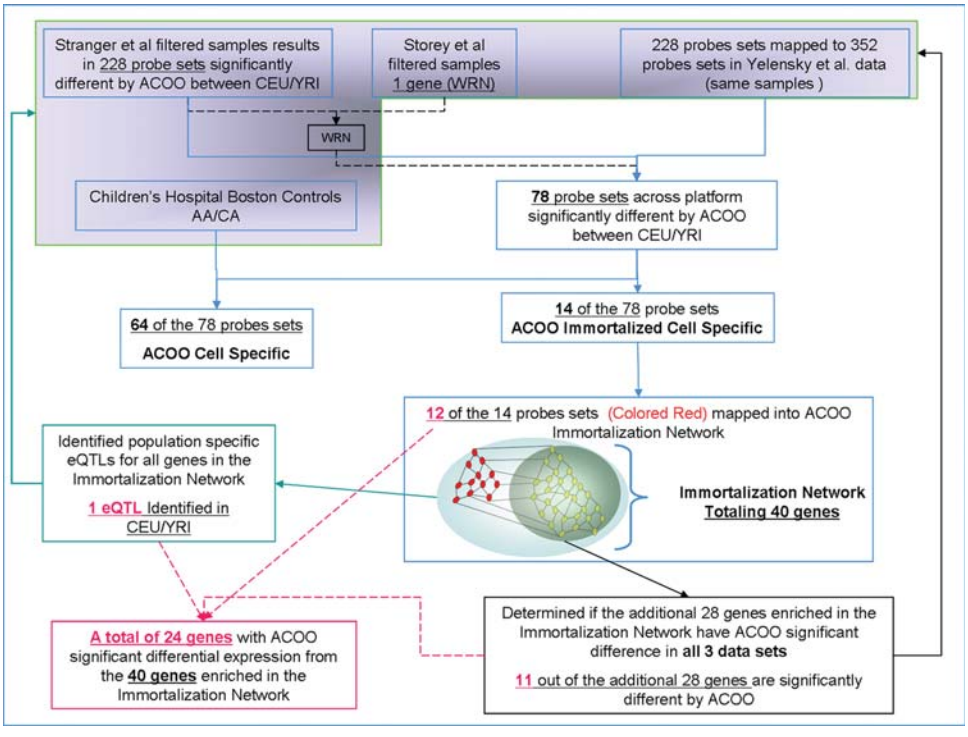

Figure 1. Analytic flow of the expression analysis of ACOO. Shaded boxes at the top represent independent data sets of gene expression profiling. The topmost three boxes are three experiments by different investigators on two expression profiling platforms measuring expression in the immortalized lymphoblasts of the YRI and CEU HapMap individuals. The fourth data set is measured on a group of children (CA and AA) who served as controls in an unrelated (autism) study. These cells in this population were not immortalized prior to measurement. Eighty probe sets were measured as significantly differentially expressed across the three immortalized cell data sets. Of those, 66 were also differentially expressed in non-immortalized data set and the subsequent analysis focused on those 14 probe sets that were only differentially expressed in the immortalized cells. Twelve of those 14 probe sets were mapped to genes in IPA, and a network (dubbed the COO Immortalization Network) of 40 genes was automatically constructed. This network was then assessed against the three original expression data sets in two ways. First, one gene was identified as having a significant eQTL based on the associated HapMap SNP data. Second, additional 11 genes from the immortalization network were differentially expressed across all three data sets in addition to the original 12 found (through a much more stringent filter).

other populations with European ancestry has to date no reported predisposition or population-specific susceptibility to EBV infection. This raises the question of the degree to which the reported expression differences are due to laboratory technique, measurement platform difference, laboratory-specific variation in EBV-driven cell immortalization, or COO-specific responses to EBV infection and immortalization. To explore this question, we filtered samples and genes to accentuate population stratification between CEU and YRI trios. Our guiding principle was to select for samples and genes with the highest consistency within ACOO and the least overlap across ACOO. Our approach is outlined in Figure 1. We analyzed four independent recent studies, three of which were conducted on immortalized cell lines previously published (5,14,15), to find the reproducible differences by ACOO across two expression array platforms (Affymetrix and Illumina), and a fourth analysis was performed on an expression experiment of primary lymphoid cells from African Americans (AAs) and Caucasians (CAs) (16). Further description of the experiments, type of array platforms and genes analyzed are listed in Supplementary Material, Table S1. To reduce noise from the varied measurement platforms and laboratory-specific technique, this analysis was intentionally driven to high specificity at the cost of sensitivity (9) by the filtering process, as described. Our analysis identified an 'immortalization network' consisting of 40 genes, of which 24 genes are differentially expressed between the CEU and YRI populations. Furthermore, one of these genes, Werner syndrome helicase (WRN), is significantly correlated with EBV titer. Subsequently, we relaxed the original aggressive filtering of the data and found the large majority of the immortalization network's genes were differentially expressed across ACOO. Moreover, we identified a *cis* eQTL in gene POLR1A in the network with respect to ACOO.

RESULTS

Identification of initial COO differential expression

We started the analysis with the reproducibility of the COO-specific differences in the first study (4), across two trios (CEU and YRI) divided into four populations: HapMap parents (YRI_p and CEU_p) and separately HapMap children (YRI_c and CEU_c). We selected those genes that were expressed most consistently within the YRI and separately CEU populations, respectively, and then identified those of the intersecting set that were significantly differentially expressed. The intersection of the number of consistently expressed genes within COO across both populations differed for the parents ($n = 1043$) when compared with their children ($n = 568$). The shared set of genes that were highly consistently expressed in both parental and child populations and that also were significantly

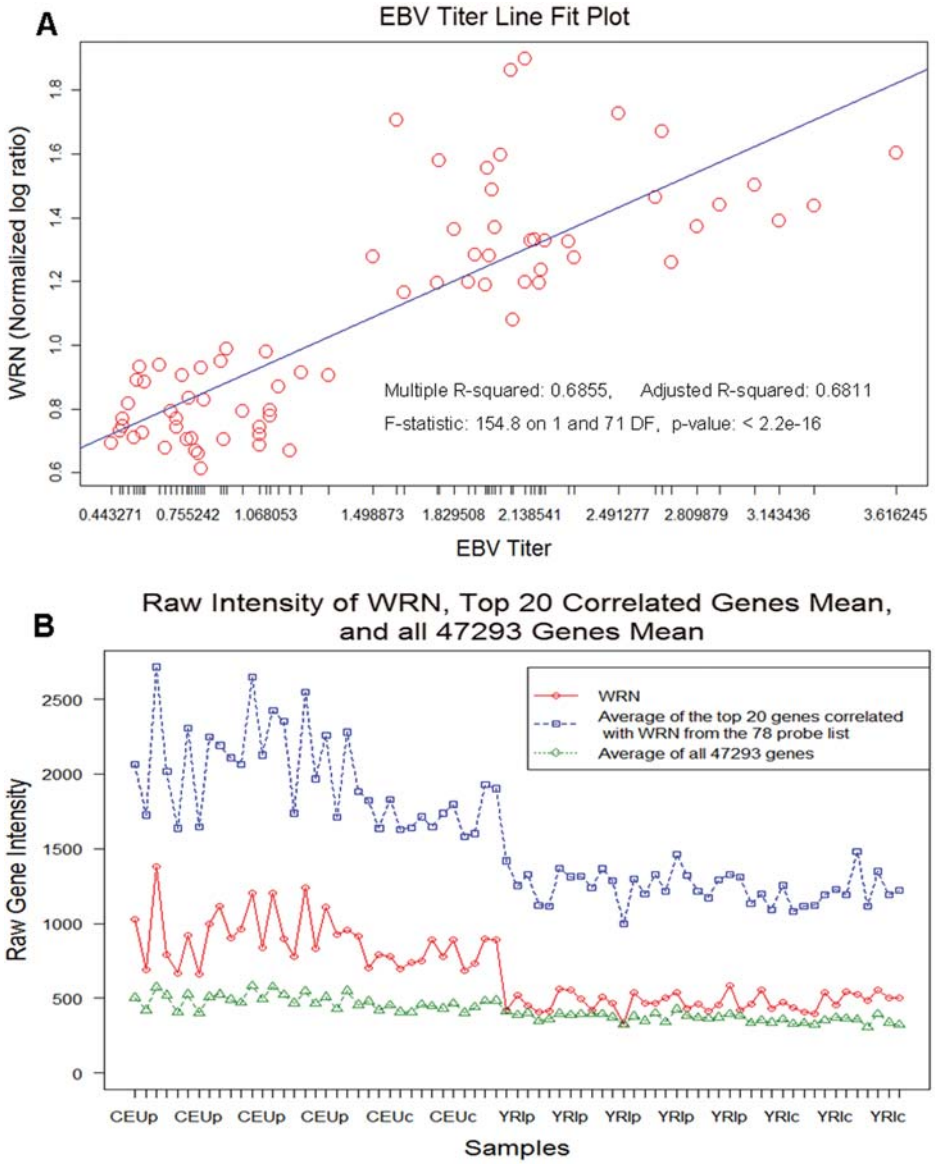

Figure 2. (**A**) Correlation of WRN to relative EBV titer across the filtered CEU and YRI samples and (**B**) the distribution of non-normalized WRN values and the mean values of the 20 genes across the CEU and YRI populations and for all the transcripts measured on the arrays.

differentially expressed after Bonferroni correction numbered 228 (Supplementary Material, Table S2). The biological functions program significantly enriched [as per the Ingenuity IPA program (17)] in the differentially expressed genes included processing and splicing of mRNA, immortalization of cells, transcription and expression of DNA, synthesis and metabolism of proteins, processing and modification of rRNA, receptor-mediated endocytosis, transport and catabolism of proteins, colony formation, activation of HIV type 1, ubiquitination and cholangiocarcinoma (data not shown). Of the 228 genes differentially expressed across ACOO, the top 20 genes most correlated with WRN, using Pearson correlation, were identified and highlighted with an '*' in Supplementary Material, Table S2. Of note, the viral titer (courtesy David Altshuler, see Materials and Methods) correlated significantly with WRN gene expression across the filtered CEU and YRI samples from Stranger *et al*. (5) with an $R^2 = 0.69$ and regression-significant $P = <2.2 \times 10^{-16}$ (Fig. 2A). Separately, the children's EBV titer correlated with WRN expression with an R^2 of 0.86 and P-value of 2.89×10^{-13}, and the parents EBV titer correlated with WRN expression with an R^2 of 0.70 and P-value of 1.18×10^{-13} (data not shown). The distribution of WRN values is much higher than the average expression of genes in the genome across all samples, which is consistent with previous reports of WRN having high levels of expression in immortalized cells. The 20 genes closely correlated with WRN also have higher mean expression across the CEU and YRI populations when compared with WRN and all the transcripts measured on the arrays (Fig. 2B).

Cross platform validation of differentially expressed genes

We conducted further analyses on an additional independent CEU and YRI population's transcriptome study. This study

was performed on the Affymetrix GeneChip Human Genome U133 Array Set HG-U133A (15). Of the 228 genes significantly different on the Illumina platform between CEU and YRI, there were 99 probe sets corresponding to the same genes significantly different on the Affymetrix platform. Of these 99 probe sets, 21 were removed because the differential expression was discordant (down for the YRI population on the Illumina platform but up regulated compared to the CEU on the Affymetrix platform) leaving 78 probe sets for further analyses (Table 1). WRN was also among the genes that were significantly different on the Affymetrix HG-U133A platform. In a third, but much smaller, data set, we applied the aforementioned filtering process on only eight CEU and eight YRI founder males from the Affymetrix Human Focus Array and only one gene, WRN, was found to be significantly different between CEU and YRI samples. That is, WRN is significantly differentially expressed in three independent studies (4,14,15). The top disease and disorders (as per the Ingenuity IPA program) enriched were viral function, connective tissue disorders (immortalization), cancer, cardiovascular disease and endocrine system disorders. WRN is among the genes in each of the top three enriched categories. The biological functions significantly enriched in the differentially expressed genes included processing and splicing of mRNA, cross-link repair of DNA, viral transactivation, immortalization of cells, transcription and expression of DNA, cell division, colony formation, contact growth inhibition, apoptosis, cell death, synthesis of proteins, gastric carcinoma (Table 2). Additionally, we performed linear regression analyses to determine the squared Pearson correlation coefficients (R^2) and *p*-values of the 20 genes most correlated with WRN (dependent variable) mRNA expression in a pairwise manner out of the 78 probe sets cross-platform validated for ACOO differential expression. We used an R^2 cutoff of 0.7. Consequently, the top 20 correlated probe sets have an R^2 between 0.69 and 0.84, and *P*-values $<2.2 \times 10^{-16}$ as described in Table 3. Sixteen (80%) of the 20 top correlated genes grouped with WRN into one biological functions network associated with *gene expression*, *infection mechanism* and *cancer* with an enrichment *P*-value of 1.0×10^{-47}. Seven of the top 20 genes are members of the final 12 gene set that comprised the immortalization network. We created an annotated network of these 20 genes entitled the *'Viral infection network'*, with the transcription factors MYC and P53 serving as the central hubs of this network (Fig. 3).

Identification of ACOO immortalization sensitive genes

To further explore which subset of the COO differentially expressed genes is specific to ACOO but not immortalization and specific to differences in the immortalization process with respect to ACOO, the results above were contrasted to an expression study of non-immortalized lymphoid cells harvested from the peripheral blood from AA and CA children. Figure 4 depicts a Venn diagram of the 78 significantly differentially expressed probe sets across platforms (Illumina and Affymetrix) between the immortalized CEU/YRI cells. Of those, 64 probe sets (82%) were confirmed to be significantly different between the AA and CA children populations. This left 14 probe sets (including WRN) that were differentially expressed across the CEU and YRI in the immortalized cell experiments.

An EBV immortalization gene network

The 14 probe sets that are significantly different between CEU and YRI immortalized cells that were not identified in non-immortalized lymphoblast cells (LCs) were mapped into Ingenuity's (IPA) package (Ingenuity® Systems, www.ingenuity.com) to determine which networks were enriched with these genes. Twelve of the 14 probe sets were mapped into IPA identifying 12 genes (two were unmapped ESTs) ARCN1, ATP5B, JMJD1B, NOL7, NUP54, PFN1, POLR2B, PRCC, PUM1, PWP1, WRN, ZNF410. The genes clustered into three significantly overrepresented/enriched networks with 10 genes mapped into the top-scoring network of DNA replication, recombination and repair with a *P*-value of 10^{-7}. JMJD18 and PUM1 mapped separately to Networks 2 and 3 . The 10 genes from Network 1 were exported into Ingenuity's Pathway editor to build a combined 'Immortalization Network' that includes JMJD18 and PUM1 (colored red in Fig. 4). There were several genes enriched in the '*Immortalization Network*' that were not part of the original 14 gene list. Subsequent to finding the marked network enrichment score, we relaxed the cutoffs in three ways, intra-population consistency criterion, *P*-value cutoff and multiple test correction (see Materials and Methods for more detail) in determining the statistical inference of the additional genes in the Immortalization Network, for the Illumina Platform only. By relaxing the aggressive filtering (of samples and genes) originally performed to increase specificity across the noisy and different expression platforms, an additional 11 genes (NUP62, BAT1, PSME3, SFRS2, PLRG1, CDC5L, EXO1, FEN1, DNAJA1, VCP and ZNF512B) were identified that have an ACOO-significant expression difference (Table 4) in the Immortalization Network (colored yellow in Fig. 4).

Continent of origin (COO) eQTLs within the associated immortalization pathway

We determined whether any of the genes in the '*Immortalization Network*' which had ACOO significant expression difference across the two immortalized and control data sets manifested heritable eQTL differences between CEU and YRI by using the public SNP data from NCBI build 36 (dbSNP b126) (http://ftp.hapmap.org/genotypes/2008-10_phaseII/). There was one gene, POLR1A (colored green in Fig. 4), with expression in the YRI cohort founders (60 samples) that associated with SNP rs12124 in a *cis* eQTL ($-\log10$ *P*-value $= 5.77 \times 10^{-9}$). POLR1A also has ACOO discordant expression across all three data sets. This eQTL finding is consistent with a previous report by Stranger *et al*. (data not shown).

DISCUSSION

The YRI is one of the native sub-Saharan populations suffering from the childhood cancer pandemic BL caused by the EBV. The International HapMap Project harvested peripheral blood

Table 1. The 78 probe sets corresponding to 53 genes drawn from the 99 probe sets list generated from the intersection between the Illumina and Affymetrix platforms of those genes that were the most consistently expressed within the YRI population and the CEU population, respectively, in both parents and in children

	RefSeq	Common	Illumina Human V6 arrays					Affymetrix U133A array		
			Illumina probe set ID	CEUp versus YRIp (MTC: Bonferroni)		CEUc versus YRIc (MTC: Bonferroni)		CEU versus YRI (MTC: Benjamini–Hochberg)		
				Fold difference (log2)	$P < 0.01$	Fold difference (log2)	$P < 0.01$	Affymetrix probe set ID	Fold difference (log2)	$P < 0.05$
1	NM_000462	UBE3A	GI_19718761-A	1.811	2.87E−46	1.399	7.06E−23	211575_s_at	1.0903921	2.75E−02
2	NM_000516	GNAS	GI_18426899-A	1.215	2.05E−13	1.174	2.87E−05	217673_x_at	1.0949674	6.97E−03
3	**NM_000553**	**WRN**	**GI_19924171-S**	**1.94**	**1.18E−50**	**1.54**	**5.25E−28**	**205667_at**	**1.1819617**	**6.51E−03**
4	NM_000938	POLR2B	GI_4505940-S	1.621	1.10E−38	1.431	3.83E−19	201803_at	1.0732888	3.46E−02
5	NM_000970	RPL6	GI_16753226-S	1.275	2.41E−19	1.255	1.20E−14	200034_s_at	1.1285841	1.90E−04
6	NM_001001973	ATP5C1	GI_4885078-S	1.369	5.61E−27	1.322	3.86E−17	214132_at	1.136883	3.77E−02
7	NM_001020658	PUM1	GI_13491165-S	1.425	5.04E−32	1.157	1.07E−06	201164_s_at	1.132332	6.73E−04
8	NM_001025105	CSNK1A1	GI_34147516-S	1.759	3.37E−53	1.41	4.87E−20	206562_s_at	1.2095745	1.88E−06
9	NM_001037637	BTF3	GI_29126237-S	1.319	1.10E−27	1.343	9.34E−18	208517_x_at	1.1164298	9.27E−04
10	NM_001037637	BTF3						211939_x_at	1.0701255	4.22E−02
11	NM_001037637	BTF3						214800_x_at	1.1117729	4.48E−02
12	NM_001253	CDC5L	GI_16357499-S	1.615	4.73E−41	1.401	1.71E−21	209055_s_at	1.2033471	3.16E−03
13	NM_001253	CDC5L						209056_s_at	1.1000586	3.90E−02
14	NM_001402	EEF1A1	GI_25453469-S	1.185	1.28E−17	1.129	1.14E−06	204892_x_at	1.0640733	4.48E−02
15	NM_001402	EEF1A1						206559_x_at	1.0533731	3.46E−02
16	NM_001402	EEF1A1						213477_x_at	1.0572026	2.50E−02
17	NM_001655	ARCN1	GI_21626463-S	1.579	4.07E−42	1.304	4.26E−14	201176_s_at	1.1239741	2.60E−03
18	NM_001686	ATP5B	GI_32189393-S	1.403	5.41E−29	1.412	1.86E−22	201322_at	1.1220461	5.21E−03
19	NM_002136	HNRNPA1	GI_4504444-A	1.43	1.26E−34	1.297	4.39E−18	201054_at	1.1059377	1.13E−03
20	NM_002136	HNRNPA1						201055_s_at	1.1102453	2.44E−02
21	NM_002136	HNRNPA1						200016_x_at	1.1229433	9.27E−04
22	NM_002136	HNRNPA1						214280_x_at	1.0547829	2.69E−02
23	NM_002136	HNRNPA1						213356_x_at	1.1725581	5.58E−03
24	NM_002568	PABPC1	GI_4505574-S	1.248	6.13E−18	1.188	1.31E−08	215157_x_at	1.0570977	3.90E−02
25	NM_002734	PRKAR1A	GI_33636720-S	1.338	2.02E−26	1.174	2.32E−07	200604_s_at	1.2528985	6.87E−03
26	NM_002799	PSMB7	GI_23110926-S	1.252	5.01E−10	1.361	9.07E−13	200786_at	1.0921669	3.01E−02
27	NM_003074	SMARCC1	GI_21237801-S	1.39	3.16E−27	1.305	3.01E−16	201072_s_at	1.2108523	6.51E−03
28	NM_003074	SMARCC1						201073_s_at	1.1169357	1.44E−02
29	NM_003074	SMARCC1						201074_at	1.1553272	1.68E−03
30	NM_003074	SMARCC1						201075_s_at	1.3044553	7.31E−04
31	NM_003079	SMARCE1	GI_21264354-S	1.514	1.96E−36	1.469	1.39E−22	211988_at	1.1161661	1.52E−02
32	NM_003079	SMARCE1						211989_at	1.1394086	5.06E−05
33	NM_003188	MAP3K7	GI_21735565-A	1.519	1.70E−40	1.283	2.45E−18	206854_s_at	1.1580838	6.36E−03
34	NM_003188	MAP3K7						211536_x_at	1.2212324	2.60E−03
35	NM_003188	MAP3K7						211537_x_at	1.2388926	1.68E−03
36	NM_003292	TPR	GI_4507658-S	1.163	2.08E−08	1.245	7.99E−12	201731_s_at	1.155994	1.33E−05
37	NM_003292	TPR						215220_s_at	1.157391	9.27E−04
38	NM_003463	PTP4A1	GI_17986281-S	1.772	1.61E−49	1.423	2.23E−19	200730_s_at	1.4767935	1.21E−05
39	NM_003463	PTP4A1						200731_s_at	1.3223011	1.32E−06
40	NM_003463	PTP4A1						200732_s_at	1.4044188	7.38E−10
41	NM_003463	PTP4A1						200733_s_at	1.1612434	2.42E−03
42	NM_003910	BUD31	GI_32171174-S	1.552	5.62E−39	1.487	1.90E−23	205690_s_at	1.2200912	3.01E−04
43	NM_004500	HNRNPC	GI_14110430-A	1.243	1.69E−18	1.317	8.75E−18	216559_x_at	1.4150078	3.95E−04
44	NM_004500	HNRNPC						221919_at	1.3094529	1.21E−05
45	NM_004500	HNRNPC						200751_s_at	1.2383838	3.26E−06
46	NM_004559	YBX1	GI_34098945-S	1.372	3.29E−29	1.206	1.88E−10	208628_s_at	1.0818504	4.22E−02
47	NM_005022	PFN1	GI_16753213-S	1.194	5.56E−14	1.245	1.92E−15	200634_at	1.1045147	4.77E−02
48	NM_005594	NACA	GI_40254826-S	1.218	2.56E−17	1.117	0.000161	208635_x_at	1.0518152	2.50E−02
49	NM_005778	RBM5	GI_5032030-S	1.288	1.78E−19	1.209	2.07E−08	209936_at	1.1065431	2.62E−02
50	NM_005791	MPHOSPH10	GI_31317304-S	1.627	2.44E−32	1.576	3.59E−27	212885_at	1.1840862	1.29E−04
51	NM_005973	PRCC	GI_40807446-S	1.568	2.08E−42	1.325	5.83E−13	208938_at	1.1030719	1.33E−02
52	NM_006627	POP4	GI_5729985-S	1.344	1.35E−21	1.313	7.84E−17	202868_s_at	1.0857801	4.23E−02
53	NM_006628	ARPP-19	GI_19923363-S	1.568	9.80E−38	1.342	3.23E−17	221483_s_at	1.1252328	1.47E−02
54	NM_006766	MYST3	GI_5803097-S	1.196	1.23E−10	1.327	1.76E−15	216361_s_at	1.0696214	2.00E−02
55	NM_006805	HNRPA0	GI_14110425-S	1.293	2.06E−13	1.373	1.90E−14	212626_x_at	1.1094681	1.74E−02
56	NM_006805	HNRPA0						214737_x_at	1.1395135	3.40E−02
57	NM_006838	METAP2	GI_27597083-S	1.582	2.78E−36	1.364	7.26E−22	213899_at	1.1602293	2.20E−03
58	NM_007062	PWP1	GI_5902033-S	1.533	2.72E−36	1.411	1.30E−21	201608_s_at	1.1194986	4.62E−03
59	NM_007363	NONO	GI_34932413-S	1.192	1.27E−11	1.218	1.28E−15	208698_s_at	1.2085309	4.91E−03
60	NM_007363	NONO						210470_x_at	1.1560814	1.66E−02
61	NM_012245	SNW1	GI_18860912-S	1.371	2.14E−25	1.334	9.62E−16	215424_s_at	1.168398	7.76E−03
62	NM_014077	FAM32A	GI_7661695-S	1.315	3.44E−24	1.24	2.80E−12	201863_at	1.0926882	2.40E−03

Continued

Table 1. *Continued*

	RefSeq	Common	Illumina Human V6 arrays					Affymetrix U133A array		
			Illumina probe set ID	CEUp versus YRIp (MTC: Bonferroni)		CEUc versus YRIc (MTC: Bonferroni)		CEU versus YRI (MTC: Benjamini–Hochberg)		
				Fold difference (log2)	$P < 0.01$	Fold difference (log2)	$P < 0.01$	Affymetrix probe set ID	Fold difference (log2)	$P < 0.05$
63	NM_014607	UBXD2	GI_34222095-S	1.254	2.64E−13	1.377	9.30E−19	212006_at	1.1622256	7.76E−03
64	NM_014607	UBXD2						212008_at	1.1967869	4.48E−02
65	NM_014691	AQR	GI_38788371-S	1.603	3.04E−40	1.573	3.89E−32	212584_at	1.1355405	1.79E−03
66	NM_014827	ZC3H11A	GI_7662231-S	1.297	3.78E−20	1.207	1.40E−08	205787_x_at	1.1662422	1.14E−04
67	NM_014827	ZC3H11A						205788_s_at	1.1195558	9.27E−04
68	NM_015138	RTF1	GI_34222098-S	1.51	1.25E−39	1.29	6.64E−18	212301_at	1.0887809	2.85E−02
69	NM_015235	CSTF2T	GI_14149674-S	1.448	1.86E−34	1.258	2.01E−12	212905_at	1.1704319	5.45E−03
70	NM_016167	NOL7	GI_15743546-S	1.485	8.80E−28	1.55	3.28E−29	202882_x_at	1.0945534	1.13E−02
71	NM_016604	JMJD1B	GI_38372908-S	1.406	3.16E−29	1.182	3.45E−09	210878_s_at	1.1330327	2.62E−02
72	NM_016648	LARP7	GI_7705400-S	1.707	2.80E−45	1.462	1.83E−28	212785_s_at	1.1426488	6.83E−03
73	NM_017426	NUP54	GI_26051236-S	1.819	1.53E−48	1.404	4.29E−21	218256_s_at	1.144816	2.16E−03
74	NM_017730	QRICH1	GI_38570096-S	1.644	1.06E−46	1.433	3.29E−23	209174_s_at	1.0828506	1.54E−02
75	NM_018011	ARGLU1	GI_8922258-S	1.418	1.05E−25	1.349	6.49E−17	218067_s_at	1.1376716	4.12E−02
76	NM_021188	ZNF410	GI_10863994-S	1.507	5.98E−40	1.233	4.13E−12	202010_s_at	1.0814468	1.49E−02
77	NM_024844	NUP85	GI_34147385-S	1.543	1.16E−29	1.498	9.33E−23	218014_at	1.2494535	6.29E−05
78	NM_052940	LRRC42	GI_31543202-S	1.75	2.18E−45	1.377	2.74E−19	215084_s_at	1.0861734	3.25E−02

WRN is given in bold.

lymphoblasts from the YRI trios and then transformed them into immortalized cells using EBV *in vitro*. This raised the question of the degree to which the previously reported expression differences are due to laboratory technique, measurement platform difference, laboratory-specific variation in EBV-driven cell immortalization or COO-specific responses to EBV infection and immortalization. To explore this question we tailored the approach outlined in Figure 1. This analysis led to the identification of an immortalization network characterizing the expression differences specific to the immortalization process of the CEU and YRI samples across three independent studies (4,14,15) and distinct from a fourth independent study of ACOO differences in non-immortalized cells of AA and CA cohorts (16). Of note, one of the genes in this network, WRN, a gene mutated in Werner Syndrome (WS), a recessive genetic disorder associated with a complex premature ageing phenotype, has been shown to modulate the efficiency of EBV immortalization of LC lines (18,19), possibly through its role in the stabilization of telomeres and telomerase and the immortalized genome (20,21). Likewise, the expression of WRN (and the other genes in the immortalization network) is highly correlated with EBV titer (Fig. 2). Sixteen (80%) of the top 20 genes most correlated with WRN and sixteen (80%) of its twenty most correlated genes grouped into one biological functions network associated with gene expression, infection mechanism and cancer here termed *'Viral infection network'*. Seven of the top 20 genes of the viral infection network are part of the final 12 genes that framed the immortalization network. At the center of this network are transcription factors MYC and P53. The MYC gene recently reported by Faumant *et al*. (22) was to be one of the two 'master' transcriptional systems activated in latency III program of EBV immortalization of B-cells. Among their reported major players in the EBV immortalization process are EXO1 and FEN1 which both directly bind to WRN and are significantly different and enriched in our reported immortalization network. In addition, p53 is among the genes in the viral infection network and was reported recently by Yi *et al*. (23) to have its transcriptional and apoptotic activities modulated by the EBV protein EBNA3C latent antigen essential for *in vitro* B-cell immortalization. This analysis does not rule out the possibility that all the observed COO differences are a function of a batch effect of the different times, techniques and laboratories involved in the immortalization process of the different HapMap populations even with observed differences in three sets of experiments. However, POLR1A's significantly up-regulated expression and the specific eQTL within the YRI founders may play a role in this population's increased sensitivity to EBV infection. Albeit circumstantial evidence, recently published by Michiels *et al*. (24) supports a possible role of POLR1A as a marker for head-and-neck cancers. Additionally, research by Shiratori *et al*. (25) reported that in WS fibroblasts, the WRN gene promotes rRNA transcription as a component of an RNA polymerase I (RPI)-associated complex, of which POLR1A is one of the core subunits (26). The Shiratori *et al*.'s study identified decreased levels of rRNA transcription compared with wild-type cells as a measurable marker for characterizing the premature aging of WS. They further showed how fibroblast cells in the presence of wild-type WRN increased rRNA levels and cell proliferation. Although further studies are required to elucidate POLR1A's role in EBV-transformed B-cell, our findings shed light on POLR1A as a component of the EBV *in vitro* cell immortalization process with a possible ACOO hereditary signature. The findings presented here are consistent with the yet unproven hypothesis that these *in vitro* results echo population health; that is, lymphoblastoid cell lines sensitivity to EBV immortalization may mirror the EBV infection pandemic in Central Africa. The aforementioned data are presented as initial evidence of a set of genes that differ in expression by ACOO and among them a subset of genes that is environmentally

Table 2. The 51 Functions identified by the IPA package for the cross-platform 78 Probes sets differentially expressed between CEU and YRI Trios

	Function	Function annotation	*P*-value	Molecules
1	Processing	processing of mRNA	2.70E−06	CDC5L, CSTF2T, HNRNPA0, NONO, PABPC1, SNW1
2	**Cross-link repair**	**cross-link repair of DNA**	**1.26E−04**	**CDC5L, WRN (includes EG:7486)**
3	Biosynthesis	biosynthesis of ADP	1.88E−04	ATP5B, ATP5C1
4	Binding	binding of Gal4p binding site	3.49E−04	SMARCC1, SMARCE1
5	Packaging	packaging of DNA	4.48E−04	MYST3, SMARCC1
6	Disruption	disruption of nucleosomes	6.81E−04	SMARCC1, SMARCE1
7	**Transactivation**	**transactivation of HIV-1**	**6.81E−04**	**SNW1, WRN (includes EG:7486)**
8	**Transcription**	**transcription**	**8.08E−04**	**BTF3 (includes EG:689), BUD31 (includes EG:8896), CDC5L, MAP3K7, MYST3, PFN1, POLR2B, PRKAR1A, PWP1, RPL6, SMARCC1, SMARCE1, SNW1, WRN (includes EG:7486), YBX1**
9	Modification	modification of RNA	8.26E−04	HNRNPC, NONO, PABPC1, RBM5
10	Expression	expression of DNA	8.99E−04	BTF3 (includes EG:689), BUD31 (includes EG:8896), CDC5L, MYST3, PFN1, POLR2B, PRKAR1A, RPL6, SMARCC1, SMARCE1, SNW1, YBX1
11	**Immortalization**	**immortalization of fibroblast cell lines**	**9.62E−04**	**PRKAR1A, WRN (includes EG:7486)**
12	**Transactivation**	**transactivation of HIV**	**1.29E−03**	**SNW1, WRN (includes EG:7486)**
13	Catabolism	catabolism of ATP	1.66E−03	ATP5B, ATP5C1
14	Polyadenylation	polyadenylation of mRNA	1.66E−03	CSTF2T, PABPC1
15	Cell division process	cell division process of oocytes	2.31E−03	GNAS, TPR
16	Cytostasis	cytostasis of cell lines	2.66E−03	METAP2 (includes EG:10988), PRKAR1A, SMARCE1, UBE3A
17	Moiety attachment	moiety attachment of mRNA	2.79E−03	CSTF2T, PABPC1
18	Contact growth inhibition	contact growth inhibition of cell lines	2.94E−03	METAP2 (includes EG:10988), PRKAR1A, SMARCE1
19	Cell division process	cell division process of female germ cells	3.60E−03	GNAS, TPR
20	Metabolism	metabolism of ATP	4.19E−03	ATP5B, ATP5C1
21	Cell death	cell death of tumor cell lines	4.42E−03	CSNK1A1, GNAS, HNRNPA1, HNRNPC, MAP3K7, PRKAR1A, RBM5, SMARCC1, SMARCE1, YBX1
22	Metabolic process	metabolic process of ATP	4.50E−03	ATP5B, ATP5C1
23	Contact growth inhibition	contact growth inhibition of eukaryotic cells	5.05E−03	METAP2 (includes EG:10988), PRKAR1A, SMARCE1
24	Apoptosis	apoptosis of tumor cell lines	5.73E−03	CSNK1A1, HNRNPA1, HNRNPC, MAP3K7, PRKAR1A, RBM5, SMARCC1, SMARCE1, YBX1
25	Processing	processing of RNA	5.81E−03	HNRNPC, NONO, RBM5
26	Contact growth inhibition	contact growth inhibition	6.81E−03	METAP2 (includes EG:10988), PRKAR1A, SMARCE1
27	Cell death	cell death of cell lines	9.49E−03	CSNK1A1, EEF1A1, GNAS, HNRNPA1, HNRNPC, MAP3K7, PRKAR1A, RBM5, SMARCC1, SMARCE1, YBX1
28	Cell division process	cell division process of germ cells	9.94E−03	GNAS, TPR
29	**Activation**	**activation of HIV-1**	**1.14E−02**	**SNW1, WRN (includes EG:7486)**
30	Cell division process	cell division process of gonadal cells	1.14E−02	GNAS, TPR
31	**Transactivation**	**transactivation of Retroviridae**	**1.18E−02**	**SNW1, WRN (includes EG:7486)**
32	Contact growth inhibition	contact growth inhibition of tumor cell lines	1.23E−02	PRKAR1A, SMARCE1
33	**Immortalization**	**immortalization of cells**	**1.28E−02**	**PRKAR1A, WRN (includes EG:7486)**
34	Splicing	splicing of mRNA	1.34E−02	CDC5L, SNW1
35	**Transcription**	**transcription of gene**	**1.94E−02**	**BTF3 (includes EG:689), MAP3K7, MYST3, POLR2B, WRN (includes EG:7486)**
36	**Transactivation**	**transactivation of virus**	**2.02E−02**	**SNW1, WRN (includes EG:7486)**
37	Papillary carcinoma	papillary carcinoma	2.21E−02	PRKAR1A, TPR
38	Development	development of animal	2.33E−02	GNAS, HNRNPC, MYST3, PRKAR1A, YBX1
39	**Expression**	**expression of gene**	**2.35E−02**	**BTF3 (includes EG:689), MAP3K7, MYST3, POLR2B, WRN (includes EG:7486)**
40	Developmental process	developmental process of leukemia cell lines	2.56E−02	JMJD1B, MYST3, PRKAR1A
41	Splicing	splicing of RNA	2.69E−02	HNRNPC, NONO
42	**Colony formation**	**colony formation of fibroblast cell lines**	**2.90E−02**	**PRCC, WRN (includes EG:7486)**
43	Developmental process	developmental process of animal	2.90E−02	GNAS, HNRNPC, METAP2 (includes EG:10988), MYST3, PRKAR1A, YBX1
44	Bipolar affective disorder	bipolar affective disorder	2.98E−02	ATP5C1, GNAS, PRKAR1A
45	**Developmental process**	**developmental process of organism**	**3.17E−02**	**GNAS, HNRNPC, METAP2 (includes EG:10988), MYST3, PRKAR1A, WRN (includes EG:7486), YBX1**
46	**Apoptosis**	**apoptosis of eukaryotic cells**	**3.53E−02**	**CSNK1A1, GNAS, HNRNPA1, HNRNPC, MAP3K7, PRKAR1A, RBM5, SMARCC1, SMARCE1, WRN (includes EG:7486), YBX1**

Continued

Table 2. *Continued*

	Function	Function annotation	*P*-value	Molecules
47	**Primary tumor**	**primary tumor**	**3.97E−02**	**EEF1A1, GNAS, HNRNPA1, LARP7 (includes EG:51574), METAP2 (includes EG:10988), PFN1, PRCC, PRKAR1A, TPR, UBE3A, WRN (includes EG:7486)**
48	Synthesis	synthesis of protein	4.21E−02	EEF1A1, METAP2 (includes EG:10988), NACA, RPL6
49	Tumorigenesis	tumorigenesis of tumor cell lines	4.75E−02	GNAS, PRKAR1A
50	**Cell death**	**cell death of eukaryotic cells**	**4.77E−02**	**CSNK1A1, EEF1A1, GNAS, HNRNPA1, HNRNPC, MAP3K7, PRKAR1A, RBM5, SMARCC1, SMARCE1, WRN (includes EG:7486), YBX1**
51	Gastric carcinoma	gastric carcinoma	4.84E−02	LARP7 (includes EG:51574), PRCC

Bold values indicate biological functions with WRN enrichment.

Table 3. The top 20 Pearson correlation coefficients (R^2), *F*-statistic and *P*-values of WRN (dependent variable) mRNA expression in a pairwise manner to all 78 probe sets cross-platform validated with ACOO differential expression

Illumina Human V6 arrays			1 and 259 degrees of freedom				
	Illumina probe set ID	RefSeq	Common	Multiple R^2	Adjusted R^2	*F*-statistic	*P*-value ($\times 10^{-16}$)
1	GI_19924171-S	NM_000553	**WRN**	1.000	1.000	5.15E + 33	<2.2
2	GI_19718761-A	NM_000462	UBE3A	0.837	0.836	1329	<2.2
3	GI_38570096-S	NM_017730	QRICH1	0.809	0.808	1097	<2.2
4	GI_7705400-S	NM_016648	LARP7	0.790	0.790	976	<2.2
5	GI_34147516-S	NM_001025105	CSNK1A1	0.782	0.781	927.1	<2.2
6	GI_21735565-A	NM_003188	MAP3K7	0.780	0.779	915.8	<2.2
7	GI_26051236-S	NM_017426	**NUP54**	0.767	0.766	853.7	<2.2
8	GI_27597083-S	NM_006838	METAP2	0.763	0.763	835.6	<2.2
9	GI_34222098-S	NM_015138	RTF1	0.763	0.762	833.3	<2.2
10	GI_31543202-S	NM_052940	LRRC42	0.761	0.760	826.3	<2.2
11	GI_5902033-S	NM_007062	**PWP1**	0.731	0.730	704.2	<2.2
12	GI_38788371-S	NM_014691	AQR	0.724	0.723	678.4	<2.2
13	GI_21626463-S	NM_001655	**ARCN1**	0.721	0.720	669	<2.2
14	GI_16357499-S	NM_001253	**CDC5L**	0.719	0.718	661.8	<2.2
15	GI_10863994-S	NM_021188	**ZNF410**	0.715	0.714	648.4	<2.2
16	GI_4505940-S	NM_000938	**POLR2B**	0.708	0.707	62808	<2.2
17	GI_31317304-S	NM_005791	MPHOSPH10	0.705	0.704	618.4	<2.2
18	GI_13491165-S	NM_001020658	**PUM1**	0.696	0.695	593.5	<2.2
19	GI_17986281-S	NM_003463	PTP4A1	0.695	0.694	590.9	<2.2
20	GI_19923363-S	NM_006628	ARPP-19	0.690	0.689	577.2	<2.2
21	GI_8922258-S	NM_018011	ARGLU1	0.690	0.689	57607	<2.2

The eight boldfaced genes are part of the final immortalization network.

sensitive to EBV in healthy individuals. Further studies are required to evaluate this hypothesis and measurements in individuals with different COO during *in vivo* EBV infection might be illuminating in this regard.

MATERIALS AND METHODS

Normalization

In the initial analysis of the Illumina Human V6 arrays used by Stranger *et al*. (4) and the Affymetrix Human Focus arrays used by Storey *et al*. (14), array probe set intensities that were <0.01 were set to 0.01. For each individual array, all probe sets were divided by the 50th percentile of all probes sets on that array and then each gene was divided by the median of its measurements across all arrays. For the U133 Array Set HG-U133A and the HG-U133-Plus-2 arrays, we applied GCRMA normalization. The expression arrays used to determine eQTLs were normalized as described in the Bioconductor program (27) GGtools 3.0 created by Vince Carey (28).

Noise reduction in Stranger *et al*.'s data set

We intentionally pursued a highly conservative analysis to maximize specificity. Each population was filtered to include only genes that have a 100% detection rate across all *in-vitro* transcriptions (IVTs) to be compared. For the first data set (4): out of the 47 293 probe sets on each array [compared between the CEU (60 samples) and YRI (60 samples) parents and children (30 samples each) groups], only 4640 probes for CEU_p and YRI_p and 4839 probes for CEU_c and YRI_c populations were detected at 100% across all IVTs. To determine the IVT replication outliers, principal component analysis of the 100% detected gene list was used. An outlier was defined as any IVT that was not within the same quarter as

Table 4. The 24 immortalization probe set Ids and ACOO expression differences in *P*-values for SekWon *et al*.'s primary LBC data set (Affymetrix.GeneChip.HG-U133_Plus_2), Yelensky *et al*.'s (Affymetrix.GeneChip.HG-U133A) and Stranger *et al*.'s (Illumina WGA-6) immortalized LBC data sets

Gene symbol	EBV immortalized B cells Illumina V-6 (CEU/YRI)			AHG-U133A (CEU/YRI)			Non-immortalized B cells HG-U133_Plus_2 (CA/AA)	
	Gene/probe ID	Fold change	*P*-value (<0.05)	Gene/probe ID	Fold change	*P*-value (<0.05)	Fold change	*P*-value (<0.05)
ARCN1	GI_21626463-S	1.5	5.84E−55	201176_s_at	1.1	1.67E−03	No significant difference	
ATP5B	GI_32189393-S	1.4	1.38E−48	201322_at	1.1	2.77E−03	No significant difference	
BAT1	GI_45580710-I	1.2	3.13E−21	200041_s_at	1.2	1.27E−05	1.5	1.67E−03
				212384_at	1.2	2.10E−02	1.9	1.66E−03
CDC5L	GI_16357499-S	1.5	2.48E−61	209055_s_at	1.2	1.92E−03	2.3	1.83E−05
				209056_s_at	1.1	2.60E−02	1.8	1.74E−03
DNAJA1	GI_4504510-S	1.8	4.77E−55	200880_at	1.1	1.67E−02	2.4	2.68E−04
				200881_s_at	1.2	2.91E−04	4.4	3.11E−10
EXO1	GI_39995067-I	1.04	4.04E−05	204603_at	1.2	2.57E−05	No significant difference	
	GI_39995068-A	1.8	1.04E−59					
FEN1	GI_19718776-S	1.8	1.64E−55	204768_s_at	1.4	2.20E−04	1.3	5.06E−03
JMJD1B	GI_38372908-S	1.3	3.65E−37	210878_s_at		1.69E−02	No significant difference	
NOL7	GI_15743546-S	1.5	8.37E−55	202882_x_at	1.1	6.93E−03	1.3	2.06E−02
NUP54	GI_26051236-S	1.6	6.19E−65	218256_s_at	1.1	1.22E−03	No significant difference	
NUP62	GI_34335245-A	1.4	4.41E−33	207740_s_at	1.2	2.74E−02	1.3	2.65E−02
PFN1	GI_16753213-S	1.2	5.76E−25	200634_at	1.1	3.06E−02	No significant difference	
PLRG1	GI_4505894-S	1.5	3.68E−53	225194_at	No Probe		1.4	2.06E−02
POLR1A[a]	GI_7661685-S	1.3	1.66E−09	222704_at	No Probe		1.2	1.34E−03
POLR2B	GI_4505940-S	1.5	3.80E−58	201803_at	1.1	2.25E−02	No significant difference	
PRCC	GI_40807446-S	1.5	1.14E−54	208938_at	1.1	7.89E−03	No significant difference	
PSME3	GI_30410793-A	1.8	2.07E−45	200987_x_at	1.2	3.51E−03	1.6	2.94E−05
				209853_s_at	1.2	2.51E−03	1.6	1.13E−05
PUM1	GI_13491165-S	1.3	1.10E−37	201164_s_at	1.1	3.39E−04	No significant difference	
PWP1	GI_5902033-S	1.5	5.68E−58	201608_s_at	1.1	2.51E−03	No significant difference	
SFRS2	GI_4506898-S	1.9	1.45E−62	200753_x_at	1.2	1.22E−03	1.4	2.06E−02
VCP	GI_7669552-S	1.3	1.78E−17	208649_s_at	1.2	3.47E−04	1.3	2.12E−02
WRN	**GI_19924171-S**	**1.8**	**1.36E−73**	**205667_at**	**1.2**	**3.53E−03**	**No significant difference**	
ZNF410	GI_10863994-S	1.4	1.67E−51	202010_s_at	1.1	9.10E−03	No significant difference	
ZNF512B	GI_34013527-S	No significant difference		55872_at	No significant difference		1.2	3.86E−02

Bold values indicate WRN gene significant difference across all three platforms.
[a]Gene with population specific eQTL.

the other replicates in the four quarters from PC1 (*x*-axis) and PC2 (*y*-axis) (Supplementary Material, Fig. S1). There had to be at least three IVTs grouped for each cell line for inclusion in the analysis. The gene intensity variation across replicated IVTs within a population was filtered to include only those probes sets with a ± 0.5 standard deviation of the mean. This resulted in the following sets of *population-consistent* probe sets: YRI_p 3121 probe sets, CEU_p 2759 probe sets, YRI_c 1640 probe sets and CEU_c with 1520 probe sets whose combined expression ranges were within a one standard deviation band spanning the population mean. Differentially expressed probe sets were identified using one-way ANOVA (false discovery rate of 0.01, *t*-test with unequal variance and Bonferroni correction for multiple testing). We then obtained the intersection of the population-consistent probe sets across YRI_p and CEU_p identifying 1043 such probe sets. We compared the mean expression of the 1043 probe sets between CEU_p and YRI_p (*t*-test with *P*-value = 0.01 and Bonferroni correction), resulting in 958 probe sets that were significantly different between CEU_p and YRI_p populations. Within the CEU_c versus YRI_c populations, there were 607 shared probe sets that were population consistent in their respective populations. We compared the mean expression differences of 607 probe sets between CEU_c and YRI_c using *t*-test as previously described; this resulted in 568 probe sets that were significantly different between CEU_c and YRI_c populations. Of the above 958 and 568 differentially expressed probes, 228 probe sets were differentially expressed in both parent and child populations. When the same analysis was performed applying the same rigorous filtering on a smaller data set of eight CEU and eight YRI founder males, the only gene differentially expressed was WRN on the Affymetrix Human Focus Array (14).

The 228 probe sets' network analysis

We used the Ingenuity Pathways Analysis program (IPA—Ingenuity® Systems, www.ingenuity.com) to analyze the set of differentially expressed probe sets. Of the 228 probe sets, we exclude 11 expressed sequence tags (ESTs), and the remaining 217 probe sets were mapped into IPA with 140 of the 217 probe sets specifically mapping into the functions/pathways by RefSeq accession numbers. With removal of redundant gene symbols, 101 genes in total enriched 269 functions and diseases annotations (FAs). Of the 269 FAs significantly enriched within the 228 probe list, we removed 237 enriched FAs that had less than three genes, *P*-values >0.05 and/or redundant names, resulting in a final 32 FA categories enriched in the differentially expressed gene list comparing CEU and YRI samples. The 32 enriched FAs are comprised

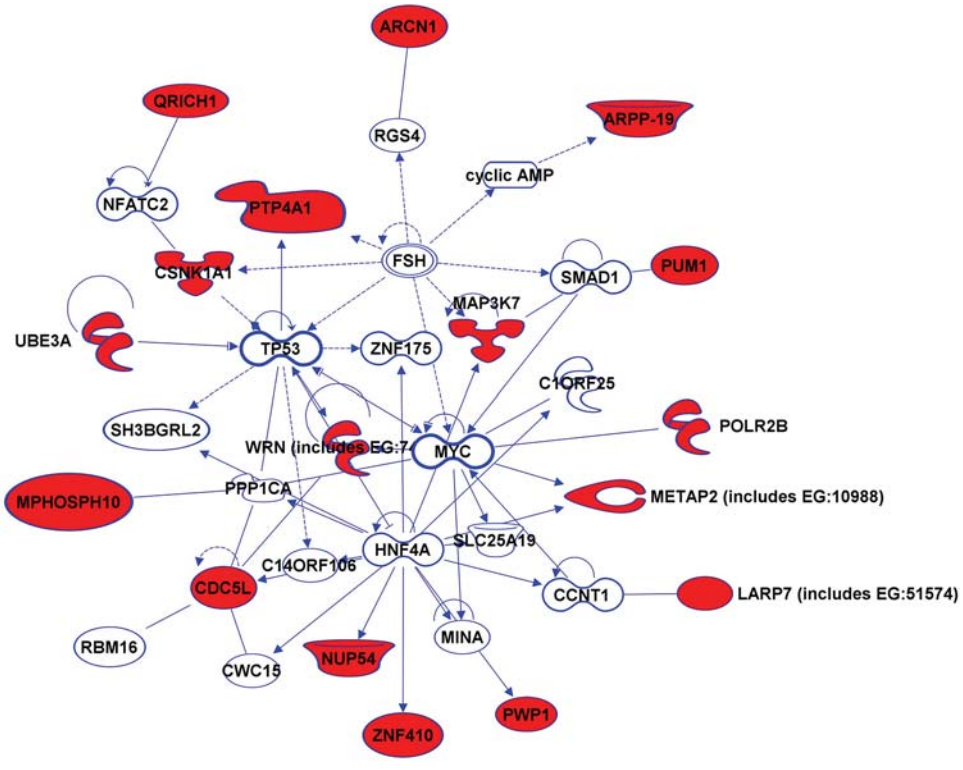

Figure 3. Of the 78 probe sets cross-platform validated with ACOO differential expression, 16 (80%) of the top 20 WRN-correlated genes (R^2 between 0.69 and 0.84) grouped with WRN into one biological functions network associated *with gene expression*, *infection mechanism* and *cancer* with an enrichment *P*-value of 1.0×10^{-47}.

of 87 (86%) of the overall 101 genes annotated in FAs by the IPA package (Data not shown).

Viral titers

Cell-line-specific viral titers were shared with us courtesy of David Altshuler and Roman Yelensky (Broad Institute, Cambridge, MA, USA). Relative EBV copy number was determined by the difference of CT method (2) and log-transformed. EBV measurements were obtained when cell-lines were first received from the Coriell Institute in 2005.

Cross platform validation of the 228 genes in Yelensky *et al*. affymetrix data set

The 228 genes identified with COO differential expression from Stranger *et al*. samples (Illumina platform) were validated across platforms using an independent study of the same samples from the CEU and YRI populations on the Affymetrix GeneChip Human Genome U133 Array Set HG-U133A (15). The initial 228-gene list mapped to 352 probe sets on the HG-U133A array by RefSeq accession number. Of the 228 genes that were significantly different on the Illumina platform between CEU and YRI, there were 78 probe sets of the same genes that were significantly different at a *P*-value cutoff of 0.05 with Benjamini–Hochberg multiple testing correction on the Affymetrix platform. The WRN gene was also among the genes that were significantly different on the Affymetrix platform, a finding that was confirmed in a third independent study of Storey *et al*.'s data on the Affymetrix Human Focus Arrays.

Squared Pearson correlation coefficients (R^2)

We preformed a liner regression analyses to determine the squared Pearson correlation coefficients (R^2) and *P*-values of WRN (dependent variable) mRNA expression in a pairwise manner to all 78 probe sets cross-platform validated with ACOO differential expression. We reported the genes with an R^2 cutoff of 0.7 or greater (Table 3).

Intersection of the immortalized cell gene list with the non-immortalized significantly different gene list

We used an in house unpublished data set of AA and CA samples consisting of 43 male and female children from 1 to 16 years of age. These samples were collected as control samples in an unrelated study of autism spectrum disorder (ASD). LCs were isolated and RNA extracted (without EBV immortalization) and hybridized to the Affymetrix U133plus2 array. The initial 228 gene list mapped to 352 probe sets on the U133plus2 array by RefSeq accession number. Statistical inference was determined using parametric test; variance assumed unequal Student's *t*-test, *P*-value cutoff 0.05, with Benjamini-Hochberg multiple test correction. Of the 524 across platform-intersected probes, 288 probe sets had significant difference between the AA and CA cohorts. We cross array (U133Pluse2 to U133A) matched the RefSeq numbers of the 288 probes yielding 299 probes for intersection across platforms. We intersected the 299 probe sets with the across platform confirmed 78

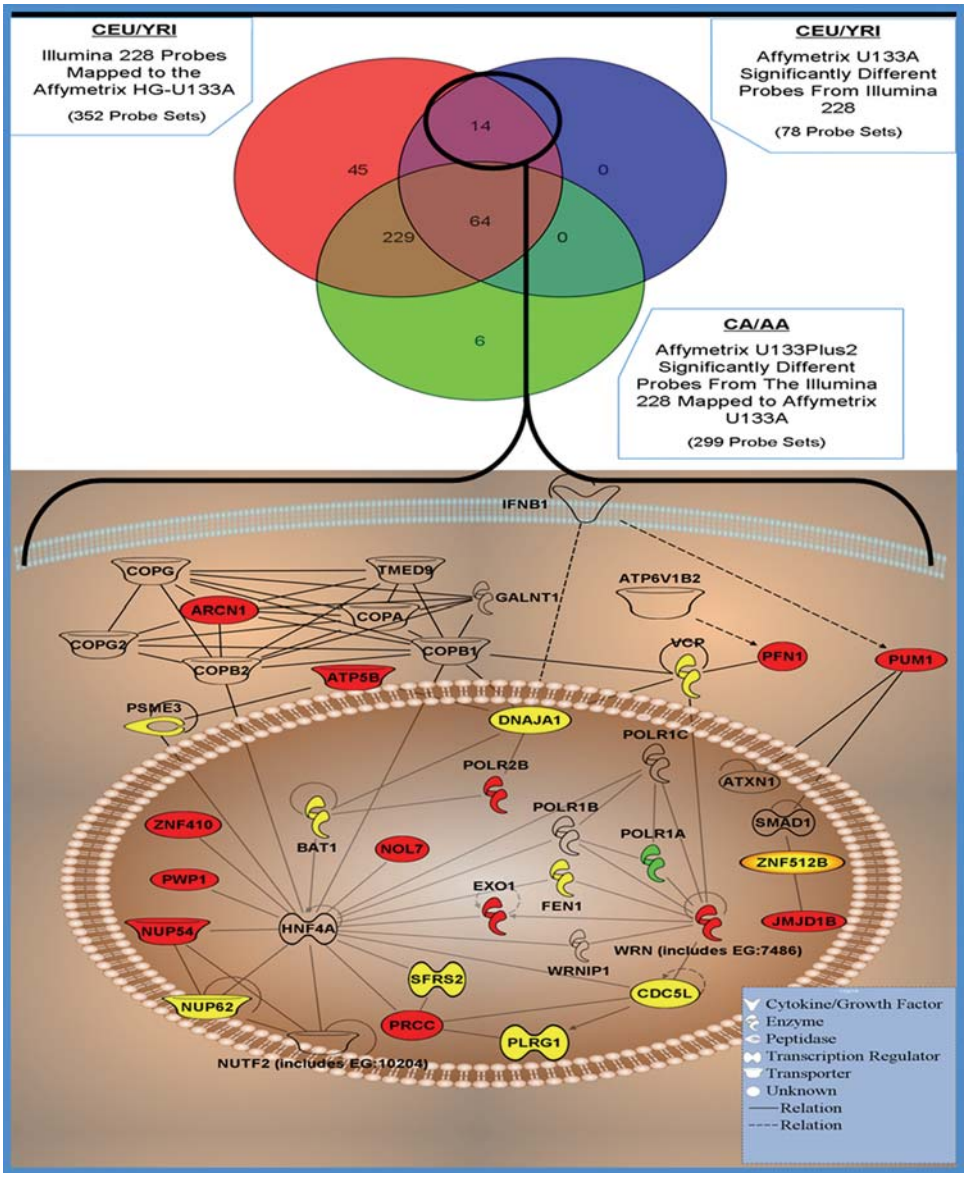

Figure 4. Twelve of the 14 probe sets identified in the Venn diagram with immortalized cell-specific differential expression (circled in Venn diagram), mapped to 12 independent genes in Ingenuity Pathway program to construct the 'immortalization network'. The 12 independent genes are depicted in red. POLR1A which has an heritable eQTL in the YRI population with significant differential expression by ACOO is in green. The additional genes with ACOO significantly different expression but are not immortalization specific are in yellow.

probe sets that have discordant expression between CEU and YRI trios.

Immortalization network enrichment

Twelve of the 14 probe sets identified as immortalized cell specific were enriched in IPA and mapped to 12 independent genes (two were unmapped ESTs). The genes clustered into 3 networks with 10 genes mapped into the top network of DNA replication, recombination and repair with a *P*-value of 10^{-27}. JMJD18 and PUM1 mapped separately to Networks 2 and 3. The 10 genes from Network 1 were exported into IPA editor to construct the '*Immortalization Network*' including JMJD18 and PUM1. To determine whether any of these additional genes have significant ACOO differential expression (subsequent to finding the marked network enrichment score), we relaxed the statistical inference cutoffs in three ways. First, we no longer filtered the genes to meet the intra-population consistency criterion. Second, we relaxed the *P*-value cutoff from 0.01 to 0.05 and, finally, we changed the multiple test correction to Benjamini–Hochberg from Bonferroni for statistical inference for the Illumina Platform only.

ACOO-specific eQTLs

The eQTLs were determined using the Bioconductor program (27) GGtools 3.0 written by Vince Carey. Here we used only

the founder population (60 parents) for the CEU and YRI cohorts. A relevant eQTL was only determined to be of interest when it was discordant for significance across the YRI and CEU populations. A significant *cis* eQTL is defined as having an SNP correlated to a gene's expression within 50 kb from the 5′ or 3′ end of the gene with a significant *P*-values less than or equal to $-\log10\ 10^{-8}$.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors are indebted to Zoltan Szallasi and Simon Kasif for critical reading and suggestions regarding biological validation. They also recognize the generous support of David Altshuler and Roman Yelenksy in providing the relative EBV viral titer data. They also thank Vincent Carey for assistance with R-GUI and Bioconductor package GGTools and GGdata, and Sek Won Kong, Christin Collins, Ingrid Holm and Lou Kunkel for providing the expression arrays of the African American and Caucasian controls from their Autism study.

Conflict of Interest statement. None declared.

FUNDING

This work was supported in part by National Library of Medicine [U54LM008748–03 to I.S.K.] and National Human Genome Research Institute [T32HG02295 to A.R.D.]. Funding to pay the Open Access publication charges for this article was provided by National Library of Medicine [U54LM008748-03].

REFERENCES

1. Allocco, D.J., Song, Q., Gibbons, G.H., Ramoni, M.F. and Kohane, I.S. (2007) Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. *BMC Genomics*, **8**, e68.
2. Echols, M.R. and Yancy, C.W. (2006) Isosorbide dinitrate–hydralazine combination therapy in African Americans with heart failure. *Vasc. Health Risk Manag.*, **2**, 423–431.
3. Jorgenson, E., Tang, H., Gadde, M., Province, M., Leppert, M., Kardia, S., Schork, N., Cooper, R., Rao, D.C., Boerwinkle, E. *et al.* (2005) Ethnicity and human genetic linkage maps. *Am. J. Hum. Genet.*, **76**, 276–290.
4. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
5. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
6. Tishkoff, S.A. and Kidd, K.K. (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.*, **36**, S21–S27.
7. Cheadle, C., Becker, K.G., Cho-Chung, Y.S., Nesterova, M., Watkins, T., Wood, W. 3rd, Prabhu, V. and Barnes, K.C. (2007) A rapid method for microarray cross platform comparisons using gene expression signatures. *Mol. Cell Probes*, **21**, 35–46.
8. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics (Oxford, England)*, **18**, 405–412.
9. Eklund, A.C. and Szallasi, Z. (2008) Correction of technical bias in clinical microarray data improves concordance with known biological information. *Genome Biol.*, **9**, R26.
10. Burkitt, D. (1958) A sarcoma involving the jaws in African children. *Br. J. Surg.*, **46**, 218–223.
11. Mutalima, N., Molyneux, E., Jaffe, H., Kamiza, S., Borgstein, E., Mkandawire, N., Liomba, G., Batumba, M., Lagos, D., Gratrix, F. *et al.* (2008) Associations between Burkitt lymphoma among children in Malawi and infection with HIV, EBV and malaria: results from a case–control study. *PLoS ONE*, **3**, e2505.
12. Ogwang, M.D., Bhatia, K., Biggar, R.J. and Mbulaiteye, S.M. (2008) Incidence and geographic distribution of endemic Burkitt lymphoma in northern Uganda revisited. *Int. J. Cancer*, **123**, 2658–2663.
13. Wakabi, W. (2008) Kenya and Uganda grapple with Burkitt lymphoma. *Lancet Oncol.*, **9**, e319.
14. Storey, J.D., Madeoy, J., Strout, J.L., Wurfel, M., Ronald, J. and Akey, J.M. (2007) Gene-expression variation within and among human populations. *Am. J. Hum. Genet.*, **80**, 502–509.
15. Choy, E., Yelensky, R., Bonakdar, S., Plenge, R.M., Saxena, R., De Jager, P.L., Shaw, S.Y., Wolfish, C.S., Slavik, J.M., Cotsapas, C. *et al.* (2008) Genetic analysis of human traits in-vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet.*, **4**, e1000287.
16. Kong, S., Collins, C., Holm, I. and Kunkel, L. (2009) *Control Samples of Autism Spectrum Disorder Hospital Program in Genomics*. Harvard Medical School, Boston, MA, USA.
17. Mayburd, A.L., Martlinez, A., Sackett, D., Liu, H., Shih, J., Tauler, J., Avis, I. and Mulshine, J.L. (2006) Ingenuity network-assisted transcription profiling: Identification of a new pharmacologic mechanism for MK886. *Clin. Cancer Res.*, **12**, 1820–1827.
18. Sugimoto, M., Tahara, H., Ide, T. and Furuichi, Y. (2004) Steps involved in immortalization and tumorigenesis in human B-lymphoblastoid cell lines transformed by Epstein-Barr virus. *Cancer Res.*, **64**, 3361–3364.
19. Sugimoto, M., Tahara, H., Okubo, M., Kobayashi, T., Goto, M., Ide, T. and Furuichi, Y. (2004) WRN gene and other genetic factors affecting immortalization of human B-lymphoblastoid cell lines transformed by Epstein–Barr virus. *Cancer Genet. Cytogenet.*, **152**, 95–100.
20. Lebel, M. and Leder, P. (1998) A deletion within the murine Werner syndrome helicase induces sensitivity to inhibitors of topoisomerase and loss of cellular proliferative capacity. *Proc. Natl Acad. Sci. USA*, **95**, 13097–13102.
21. Leder, A., Lebel, M., Zhou, F., Fontaine, K., Bishop, A. and Leder, P. (2002) Genetic interaction between the unstable v-Ha-RAS transgene (Tg.AC) and the murine Werner syndrome gene: transgene instability and tumorigenesis. *Oncogene*, **21**, 6657–6668.
22. Faumont, N., Durand-Panteix, S., Schlee, M., Gromminger, S., Schuhmacher, M., Holzel, M., Laux, G., Mailhammer, R., Rosenwald, A., Staudt, L.M. *et al.* (2009) c-Myc and Rel/NF-kappaB are the two master transcriptional systems activated in the latency III program of Epstein–Barr virus-immortalized B cells. *J. Virol.*, **83**, 5014–5027.
23. Yi, F., Saha, A., Murakami, M., Kumar, P., Knight, J.S., Cai, Q., Choudhuri, T. and Robertson, E.S. (2009) Epstein–Barr virus nuclear antigen 3C targets p53 and modulates its transcriptional and apoptotic activities. *Virology*, **388**, 236–247.
24. Michiels, S., Danoy, P., Dessen, P., Bera, A., Boulet, T., Bouchardy, C., Lathrop, M., Sarasin, A. and Benhamou, S. (2007) Polymorphism discovery in 62 DNA repair genes and haplotype associations with risks for lung and head and neck cancers. *Carcinogenesis*, **28**, 1731–1739.
25. Shiratori, M., Suzuki, T., Itoh, C., Goto, M., Furuichi, Y. and Matsumoto, T. (2002) WRN helicase accelerates the transcription of ribosomal RNA as a component of an RNA polymerase I-associated complex. *Oncogene*, **21**, 2447–2454.
26. Suzuki, N., Shimamoto, A., Imamura, O., Kuromitsu, J., Kitao, S., Goto, M. and Furuichi, Y. (1997) DNA helicase activity in Werner's syndrome gene product synthesized in a baculovirus system. *Nucleic Acids Res.*, **25**, 2973–2978.
27. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
28. Carey, V.J., Davis, A.R., Lawrence, M.F., Gentleman, R. and Raby, B.A. (2009) Data structures and algorithms for analysis of genetics of gene expression with Bioconductor: GGtools 3.x. *Bioinformatics (Oxford, UK)*, **25**, 1447–1448.