Chromosomal variation in lymphoblastoid cell lines

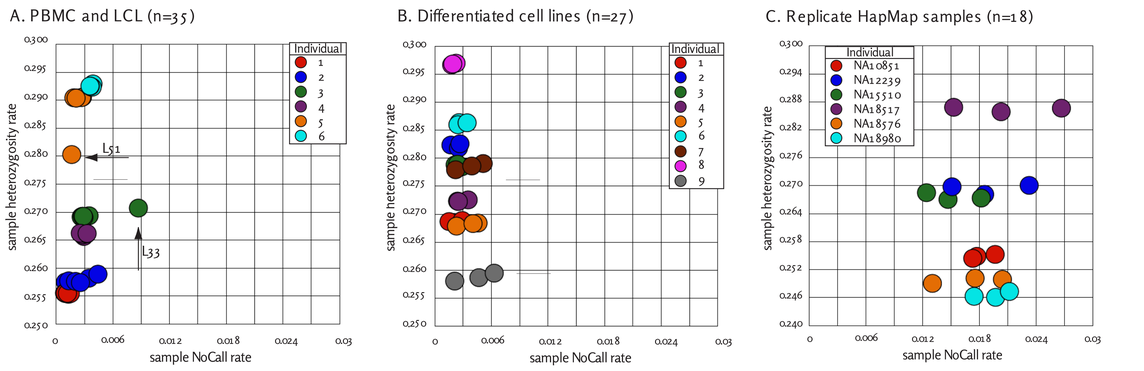
Matthew D. Shirley[1], Joseph D. Baugher[1], Eric L. Stevens[2], Zhenya Tang[3], Norman Gerry[3], Christine M. Beiswanger[3], Dorit S. Berlin[3], and Jonathan Pevsner[1,2,4,5]

[1] Program in Biochemistry, Cellular and Molecular Biology, Johns Hopkins School of Medicine, Baltimore, MD 21205
[2] Program in Human Genetics, Johns Hopkins School of Medicine, Baltimore, MD 21205
[3] Coriell Institute for Medical Research, 403 Haddon Avenue, Camden NJ 08103
[4] Department of Neurology, Hugo W. Moser Research Institute at Kennedy Krieger, Baltimore, MD 21205
[5] Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205

Abstract

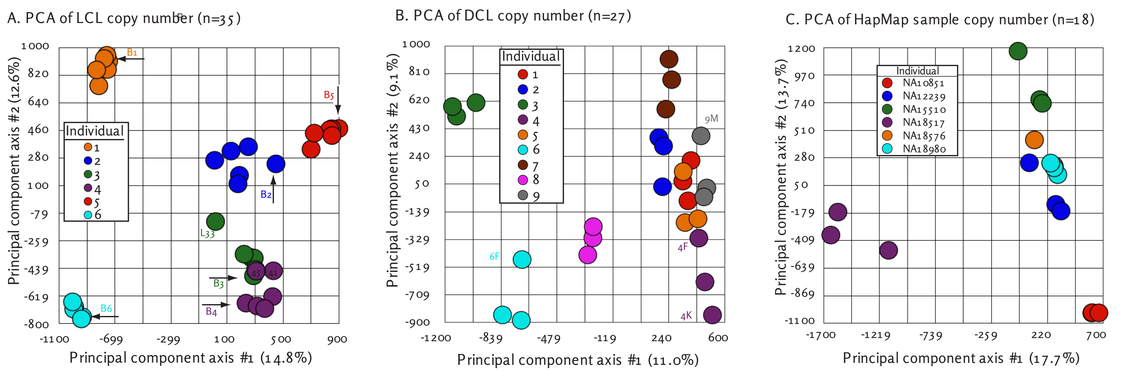
Tens of thousands of lymphoblastoid cell lines (LCLs) have been established by the research community, providing nearly unlimited source material from samples of interest. LCLs are used to address questions in population genomics, mechanisms of disease, and pharmacogenomics. Thus, it is of fundamental importance to define the extent of chromosomal variation in LCLs. We measured variation in genotype and copy number in multiple LCLs derived from peripheral blood mononuclear cells (PBMCs) of single individuals as well as two comparison groups: **(1)** three types of differentiated cell lines (DCLs) and **(2)** triplicate HapMap samples. We then validated and extended our findings using data from a large study consisting of samples from blood or LCLs. We observed high concordances between genotypes and copy number estimates within all sample groups. While the genotypes of LCLs tended to faithfully reflect the genotypes of PBMCs, 13.7% (4 of 29) of immortalized cell lines harbored mosaic regions greater than 20 megabases which were not present in PBMCs, DCLs, or HapMap replicate samples. We created a list of putative LCL-specific changes (affecting regions such as immunoglobulin loci) that is available as a community resource.

1 Genotype concordance between related samples



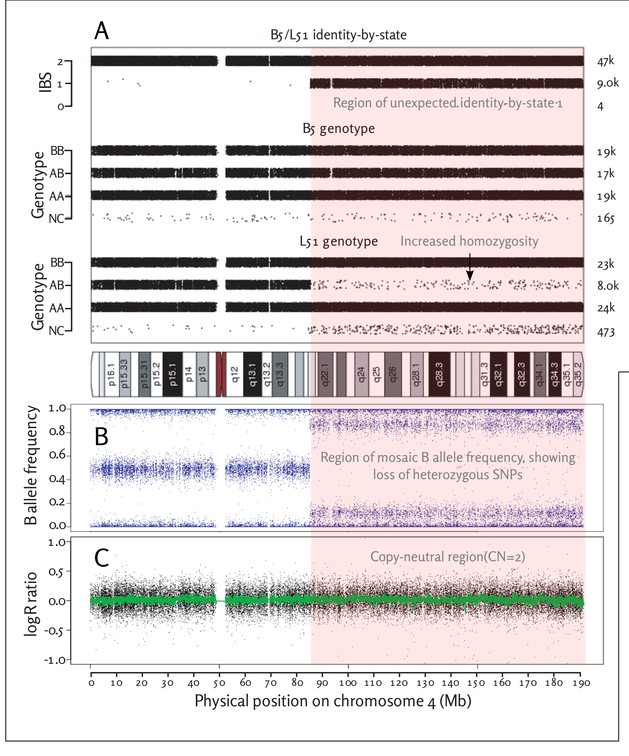
As a measure of quality-control, we measured sample heterozygosity rates (y-axis) compared to sample NoCall rates (x-axis) for **(A)** PBMC and LCL samples, **(B)** differentiated cell lines, and **(C)** HapMap replicate samples. **Several outliers are indicated, having relative differences in heterozygosity and/or NoCall rates.**

2 Principal Components Analysis (PCA) of signal intensity



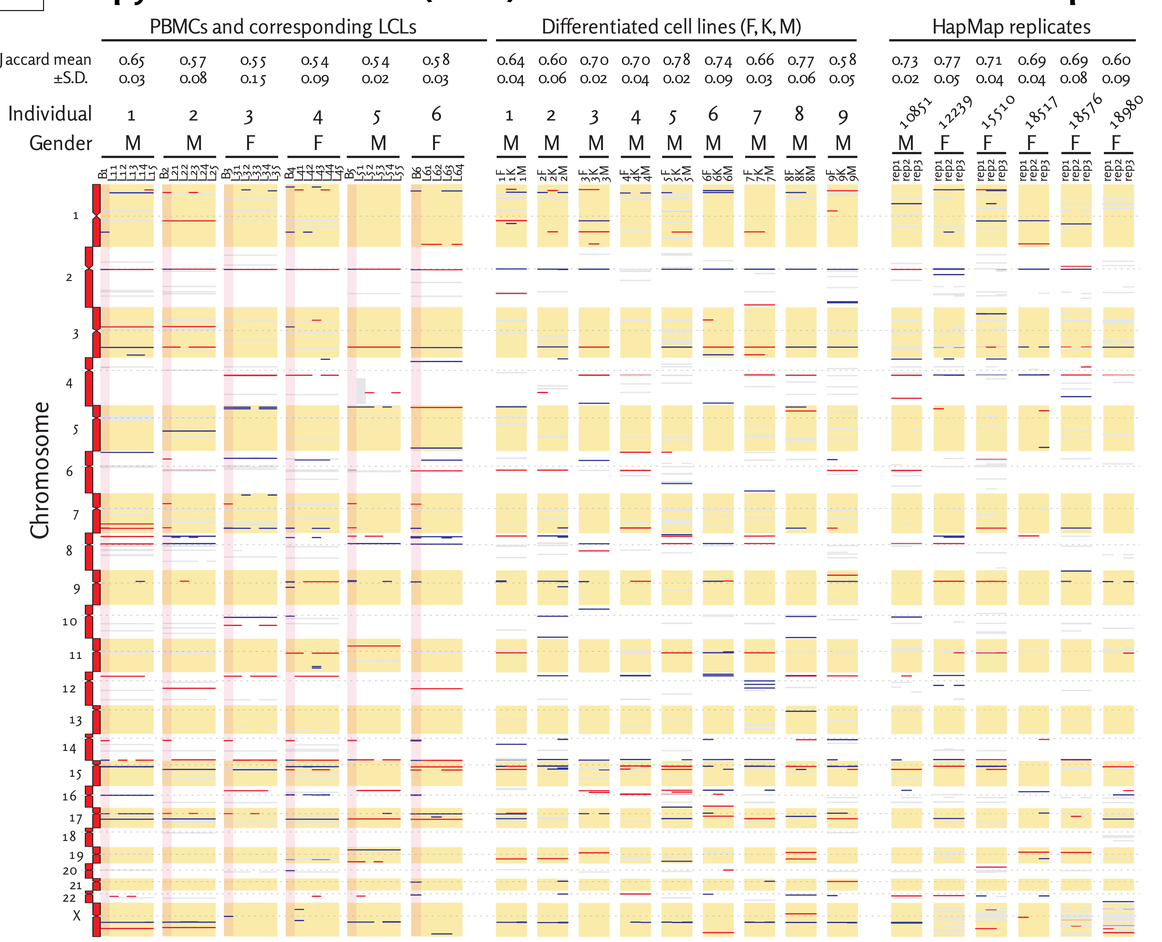
Principal components analysis of copy number data from **(A)** PBMC and LCL samples (n=35 samples derived from six individuals) **(B)** differentiated cell lines (n=27 samples derived from nine individuals), and **(C)** HapMap replicate samples (n=18 samples derived from six individuals). Values on the principal component axes correspond to percent of the variance explained. **Note that samples from the same individual cluster together, implying that there is less variance within than between samples.**

3 Mosaic Uniparental Disomy



Analysis of genotype and copy number changes in LCL individual sample L51 revealed uniparental disomy (UPD) on chromosome 4 of sample L51. (A) Plot of chromosome 4 using SNPduo software (Roberson and Pevsner, 2009). *Top panel* shows identity-by-state including a region of IBS2 (typical of replicate or identical samples) extending from 0-90 Mb followed by a region of IBS1 extending for 100 Mb (from 90 Mb to 190 Mb). *Middle panel*: genotypes for B5 showing typical patterns of BB, AB, AA, and NoCalls (NC). *Lower panel*: genotypes for L51 showing prominent decrease in heterozygous (AB) calls from 90 to 190 Mb, with an increase in the NoCall rate. **(B)** B allele frequency for sample L51 revealed a typical pattern (0-90 Mb) corresponding to BB, AB, and AA genotypes at y-axis values of 1.0, 0.5, and 0 respectively. From 90-190 Mb a shift in the pattern occurred resulting in four bands and indicated a mosaic abnormality. **(C)** logR ratio, reflecting copy number, indicated that L51 had no gain or loss of chromosomal material.

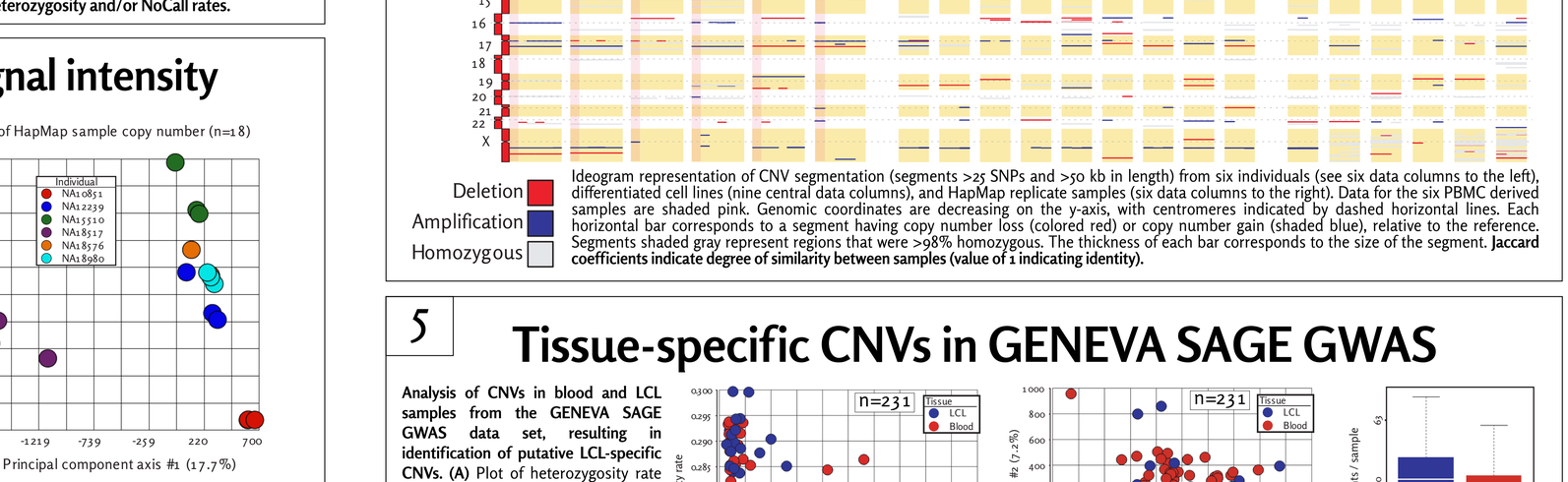
4 Copy Number Variant (CNV) concordance between related samples



Ideogram representation of CNV segmentation (segments >25 SNPs and >50 kb in length) from six individuals (see six data columns to the left), differentiated cell lines (nine central data columns), and HapMap replicate samples (six data columns to the right). Data for the six PBMC derived samples are shaded pink. Genomic coordinates are decreasing on the y-axis, with centromeres indicated by dashed horizontal lines. Each horizontal bar corresponds to a segment having copy number loss (colored red) or copy number gain (shaded blue), relative to the reference. Segments shaded gray represent regions that were >98% homozygous. The thickness of each bar corresponds to the size of the segment. **Jaccard coefficients indicate degree of similarity between samples (value of 1 indicating identity).**

5 Tissue-specific CNVs in GENEVA SAGE GWAS

Analysis of CNVs in blood and LCL samples from the GENEVA SAGE GWAS data set, resulting in identification of putative LCL-specific CNVs. (A) Plot of heterozygosity rate (y-axis) compared to sample NoCall rate (x-axis) for GENEVA SAGE replicate samples. **(B)** Principal components analysis of copy number data from GENEVA SAGE replicate samples. **(C)** Boxplot of the number of segments per sample after filtering using CNVineta software. **(D)** Manhattan plot showing association of CNV regions with blood samples or LCLs (y-axis; -log10 of probability value) as a function of chromosomal position (x-axis). Several peaks are indicated: chr6:32,066,939-32,114,701 encompassing the HLA locus, chr14:40,739,852-40,739,853 having no annotated genes, chr14:105,268,029-105,397,778 having 17 genes including immunoglobulin loci for IGHA2, IGHE, IGHG1, and IGHD, chr22:21,028,552-21,443,164 having 17 genes including immunoglobulin *lambda* loci, and chr22:21,543,587-21,570,027 having four genes including immunoglobulin *lambda* loci. **(E)** Visualization of a region showing the greatest difference between LCLs and blood samples (see **panel D**, peak at chromosome 22). The number of samples (y-axis) is plotted by chromosomal position (x-axis, chr22:20,745,308-21,609,299). Each horizontal bar in the plot corresponds to a CNV segment in a single sample from the GENEVA SAGE data set.



Conclusions

- Lymphoblastoid cell lines usually faithfully reflect the genotype and copy number of PBMCs from which they are derived.
- Occurrence of large regions of mosaic UPD and aneuploidy in 4 /29 (13.7%) LCLs suggests that it is appropriate to characterize LCLs via SNP array genotyping or other methods before performing further studies.
- We have generated a list of putative LCL-specific changes (resulting from analysis of GENEVA SAGE), which may prove useful when utilizing LCLs in genomic studies.
- Studying these same cell lines over multiple passages may allow analysis of the temporal aspect of LCL-specific genomic variation.

This work was supported by NIGMS HGCR contract HHS-N-263-2009-00026-C.