**In utero cytomegalovirus infection and development of childhood acute lymphoblastic leukemia**

Stephen Starko Francis1,2*, Amelia D. Wallace1, George A. Wendt1, Linlin Li3, Fenyong Liu4, Lee W. Riley1,3, Scott Kogan5, Kyle M. Walsh2, Adam J. de Smith2, Gary V. Dahl6, Xiaomei Ma7, Eric Delwart3, Catherine Metayer1 and Joseph L. Wiemels2

1-Department of Epidemiology, University of California, Berkeley, California, USA
2-Helen Diller Family Comprehensive Cancer Center and Departments of Neurological Surgery and Epidemiology and Biostatistics, University of California, San Francisco, California, USA
3-Blood Systems Research Institute, Department of Laboratory Medicine, University of California, San Francisco, California, USA
4-Department of Vaccinology and Infectious Diseases, University of California, Berkeley, California, USA
5-Helen Diller Family Comprehensive Cancer Center & Department of Laboratory Medicine, University of California San Francisco, San Francisco, California, USA
6-Pediatric Hematology/Oncology, Department of Pediatrics, Stanford University School of Medicine, Palo Alto, California, USA
7-Department of Chronic Disease Epidemiology, School of Public Health, Yale University, New Haven, Connecticut, USA

*Corresponding author:
Dr. Stephen Starko Francis

**Running Title:** In utero CMV infection is a risk factor for ALL

**Funding:** NIEHS R01ES09137, NHLBI R01HL105770, NCI 1T32CA151022-01, NCI R01CA155461, NCI R01CA185058

**Ethics Statement:** All data and samples have been approved for use by the University of California, Berkeley Institutional Review Board (California Childhood Leukemia Study/Protocol # 2010-10-2438 and California State Institutional Review Board (Childhood Cancer Record Linkage Project/Protocol # 12-07-0529 and California Childhood Leukemia Study/Protocol # 2010-10-2438

**Key Points:**
1- Cytomegalovirus (CMV) is prevalent in pre-diagnostic bone marrow from childhood ALL and not in AML.
2- In utero infection with CMV is a risk factor for ALL (OR=3.71, p=0.0016) and is more pronounced in Hispanics (OR=5.90, p=0.006).
Abstract:

It is widely suspected, yet controversial, that infection plays an etiologic role in the development of acute lymphoblastic leukemia (ALL), the most common childhood cancer and a disease with a confirmed prenatal origin in most cases. We investigated infections at diagnosis and then assessed the timing of infection at birth in children with ALL and age, gender, and ethnicity matched controls to identify potential causal initiating infections. Comprehensive untargeted virome and bacterial analyses of pre-treatment bone marrow specimens (n=127 ALL compared to 38 acute myeloid leukemia cases as comparison group) revealed prevalent cytomegalovirus (CMV) infection at diagnosis in childhood ALL, demonstrating active viral transcription in leukemia blasts as well as intact virions in serum. Screening of newborn blood samples revealed a significantly higher prevalence of in utero CMV infection in ALL cases (n=268) than healthy controls (n=270) (OR=3.71, CI=1.56-7.92, p=0.0016). Risk was more pronounced in Hispanics (OR=5.90, CI=1.89-25.96) vs. non-Hispanic whites (OR=2.10 CI= 0.69-7.13). This is the first study to suggest that congenital CMV infection is a risk factor for childhood ALL and is more prominent in Hispanic children. Further investigation of CMV as an etiologic agent for ALL is warranted.

Introduction:

A role for infection in the etiology of childhood acute lymphoblastic leukemia (ALL) has long been hypothesized\(^1\)\(^-\)\(^3\). While many individual infectious agents and proxies have been investigated \(^4\)\(^-\)\(^8\), a comprehensive untargeted study of possible etiologic viruses and bacteria has previously not been attempted.

Studies of tumor genetic changes in newborn blood specimens (NBS) provide unambiguous evidence of the presence of ALL initiation in utero for some subtypes of the disease\(^9\)\(^-\)\(^11\). Decreased IL-10 levels at birth in children who later develop ALL\(^12\) and
more medically-diagnosed infections early in life\textsuperscript{7,13} suggest that dysregulated immunity plays an etiologic role. Hispanics have the highest risk of ALL which appears to be increasing\textsuperscript{14}. In this study we seek to answer fundamental etiologic questions: 1) What infections are unique to childhood ALL compared to acute myeloid leukemia (AML, used here as an immune suppressed control group at the time of diagnosis). 2) Are such infections acquired \textit{in utero} and are they risk factors for ALL?

**Methods**

We conducted two complementary studies as briefly described below (Details available in Supplemental Materials).

1) \textit{Viral and Bacterial Discovery at Diagnosis}: a comprehensive next-generation sequencing-based virome and bacterial metagenomic analysis was conducted in pre-treatment diagnostic bone marrow of both childhood ALL and AML drawn from the California Childhood Leukemia Study (CCLS)\textsuperscript{15}. Two independent patient sample sets were selected from the CCLS (See Supplemental Methods) and interrogated using two separate metagenomic methods (Figure 1). The first method used 36 diagnostic bone marrow (BM) samples from children with ALL and 10 diagnostic BM samples from children with AML and were examined using virus isolation by ultra centrifugation and 454 sequencing. The second method used total RNA extracted from 91 ALL BM pooled into 4 groups and 28 AML BM pooled into 2 groups, then sequenced by Illumina Hi-Seq2500. Data generated from each method were analyzed using separate custom bioinformatics pipelines that categorized all non-human reads using BLAST (see Supplemental Methods). Additionally we used 16s based bacterial metagenomics on 91 ALL BM and 28 AML BM as a control for bacterial diversity and differential bacterial contamination of the BM samples (see Supplemental Material). Sample size was a determined based on sample availability and budget constraints.
2) **Viral Screen at birth:** NBS from 268 ALL cases and 270 cancer-free controls, randomly drawn from the California Department of Public Health (*See Supplemental Methods, Supplemental Table 2*) were screened for two herpesviruses identified in the discovery set at diagnosis (i.e., cytomegalovirus [CMV] and Epstein-Barr virus [EBV]). Viral screen was done by a droplet digital PCR-based direct detection method, specifically ultrasensitive ‘third generation’ ddPCR (Figure 1).

**Results/Discussion**

In general, ALL patients were younger than AML patients, but were otherwise demographically similar (Supplemental table 1). A limitation of these methods is the ability to primarily detect active infections, although evidence of viruses and bacteria was detected in all pretreatment bone marrow specimens from both ALL and AML groups by both Illumina sequencing (Supplemental figures 1-7) and virus enrichment isolation/454 sequencing. Only herpesviruses distinguished ALL from AML in both analysis sets; all 4 ALL Illumina pools expressed CMV transcripts, while no CMV transcripts were detected in AML pools (Figure 2A). Particle isolation and 454 sequencing showed a greater prevalence of CMV (OR=18, p= 0.003) in individual ALL bone marrow specimens as compared to AML (Figure 2B).

We initiated this study with no *a priori* candidate, following identification of CMV at diagnosis and given the *in utero* origins of ALL. Children who went on to develop ALL were 3.71 times more likely to be CMV positive at birth (p=0.0016) (Table 1). We found suggestive evidence for statistical interaction between CMV positivity and Hispanic ethnicity (logistic interaction term p=0.227). Stratification by Hispanic ethnicity showed 5.9-fold increased risk of ALL in Hispanics infected perinatally with CMV (CI=1.9-26.0). The prevalence of CMV infection was higher in whites overall and an increased risk for ALL was also demonstrated in this group; however, the result was not statistically significant (OR=2.10 CI= 0.69-7.13). We observed no difference in EBV prevalence at birth between cases and controls (11 positive cases and 11 positive controls, OR=1.01
p=1.0). After stratifying by ethnicity, no significant effect modification was observed (Table 1). CMV viral load among positive case-samples averaged 0.214 copies/ul and was higher than the CMV viral load among positive control-samples (0.071 copies/ul, Supplemental Table 2, p= 0.003). There was no significant difference in the age at diagnosis between CMV positive (mean age 4.8 years) and CMV negative (mean age 5.15) ALLs at birth.

Several studies of childhood ALL have examined herpesviruses at diagnosis with conflicting results. Two previous studies examined herpesviruses in NBS from children that developed ALL. The first examined EBV and HHV-6, where no association was found, but did not assess CMV. Another screened NBS for CMV from 48 cases and 46 controls yet found no infected children in either group. With a significantly larger sample size and a more sensitive assay, we could detect extremely low quantities of CMV DNA that may otherwise have been missed. Further replication of our findings is warranted using higher input volumes of NBS DNA, protein based detection methods, and samples from geographically diverse locations.

CMV is a common virus that infects over 90% of adults worldwide. In the United States, seroprevalence of adult CMV ranges from ~50-80% is highest in Hispanics and blacks, which is particularly interesting given the high rate of ALL in Hispanics. Timing of infection has important implications. In utero infection with CMV is a leading cause of birth defects, specifically hearing loss. While primary CMV infection during pregnancy yields the greatest risk for vertical transmission, most women of childbearing age are previously infected with CMV and therefore vertical transmission of reactivated virus accounts for the majority congenital CMV infections. Presence of CMV in the child prior to birth may have important immune control implications. Infection prior to the development of a robust adaptive immune response in the fetus and neonate may affect central tolerance and enable CMV to persist in an infectious course unlike the majority of children whom are infected after birth, thus leading to increased risk of developing ALL.
Two features of CMV infection support a role in oncogenesis in ALL. First, congenital CMV infection has been noted to cause chromosomal instability, which is suspected to be related to its teratogenic properties\(^{19}\). This is of interest in leukemia as specific chromosomal lesions are present at birth in children with ALL\(^{20}\). Second, CMV has the largest genome of any known human viral pathogen and harbors many immune evasion genes, indicating that host immune dysregulation is a critical aspect of the CMV lifecycle. The functions of many CMV genes remain poorly understood and it is likely that its genomic complexity is the direct result of the evolution of latency. Along these lines, CMV has been hypothesized to be an “oncomodulator” in adult gliomas\(^{21}\). A delicate balance has evolved between CMV and the immune system, which has led to a high population prevalence of the virus with relatively low morbidity. Nevertheless, disruptions to this balance can have serious and complex pathologic consequences for the host.

This study provides a putative candidate for future investigations. The comparison of ALLs vs. AMLs was designed solely to generate a putative agent in ALL and was not matched by race, ethnicity or age. Furthermore, the preponderance of lymphoid vs. myeloid cells in the BM samples may bias the results in relation to viruses with a tropism for B-cells. A major limitation in the DBS CMV screen is our lack of information on ALL subtype and cytogenetic characteristics. Additionally the low input quantity of DBS DNA may be resulting in false negatives, though this may drive results towards the null. Additional studies defining \textit{in utero} CMV infection in terms of ALL subtype, cytogenetics, and overall prevalence are needed.

Our findings lead us to hypothesize that \textit{in utero} or perinatal CMV infection initiates immune dysregulation during the critical period of fetal immune development, allowing a greater number of and more fulminant infections later in life as observed in epidemiologic studies\(^{7,13}\). Further validation of our findings is warranted to establish ALL as an additional reason to develop an effective CMV vaccine\(^{22}\).
Conflict-of-interest disclosure: The authors declare no competing financial interests.

Acknowledgements:
We thank the families that participate in the CCLS. Without their time and effort none of our studies would be possible. We also thank our clinical collaborators throughout California for their continued support of our research and commitment to their patients. Finally, we thank the California Department of Public Health Genetic Disease Screening Program.

Author Contributions:
Contributions: S.S.F. conceived and designed the study, performed laboratory experiments, analyzed data and wrote the primary and subsequent versions of the manuscript. C.M. and G.V.D enrolled participants and obtained samples. A.D.W. and L.L. helped to design and conduct laboratory experiments. J.A.W. provided bioinformatic and data analysis support. S.K., A.D. and K.M.W. aided in refining the study and helped prepare the manuscript. E.D., F.L. and L.W.R helped to design the study. J.L.W. helped to design and conduct the study and prepare the manuscript. All co-authors have reviewed/edited the manuscript.

References


Figures with Captions

Figure 1: Flowchart of approach
Figure 2: Non-human sequence search identification ("Blast hits") from Pre-treatment Bone Marrow. a) Results of deep sequencing RNA pools (73 ALLs in 4 pools and 28 AMLs in 2 pools) after ribosomal rRNA depletion (RiboZERO). Raw Illumina reads were aggressively quality filtered, then aligned to hg19 using BowTie2, non-aligning reads were categorized using an E-value <= 1xE-10 using blastn and the nt database. Cytomegalovirus (CMV) hits were normalized by sequencing effort per pool (Blast hits/quality reads x 1.0xE8). Odds ratios were calculated comparing presence/absence of virus in ALL vs AML pools, p-values were computed using Fisher’s exact test (for additional detail see supplemental methods). b) Results of particle isolation and 454 sequencing from an independent 36 ALL and 10 AML patients. Virus was isolated using centrifugation and Millipore filter, and then viral nucleic acids were extracted and sequenced using Roche 454 (for additional detail see supplemental methods). Recovered sequences were assembled into contigs where alignments to hg19 were removed using Bowtie2. Contigs were then categorized with a cutoff of E-value <= 1.0xE-10 using blastn and the refseq viral database. Cytomegalovirus contigs were normalized by recovered contigs per patient (Blast hits/Contigs assembled x 1xE3); odds ratios were calculated comparing presence/absence of virus in ALL vs AML patients; p-values were computed using Fisher’s exact test (for additional detail see supplemental methods). CMV was found to be the only virus that showed statistically significant variation between ALL and AML in the Illumina and 454 groups.
**Table 1: Results of Neonatal Blood Spot Screen: Cytomegalovirus (CMV) and Epstein Barr Virus (EBV)**

<table>
<thead>
<tr>
<th></th>
<th>CMV -</th>
<th>CMV +</th>
<th>OR (CI)</th>
<th>p-value</th>
<th>EBV -</th>
<th>EBV +</th>
<th>OR (CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ALL</td>
<td>242</td>
<td>26</td>
<td>3.71 (1.71-8.95)</td>
<td>0.0016</td>
<td>257</td>
<td>11</td>
<td>1.01 (0.42-2.42)</td>
<td>1</td>
</tr>
<tr>
<td>Controls</td>
<td>262</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>259</td>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>By Race</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hispanic ALL</td>
<td>134</td>
<td>15</td>
<td>5.90 (1.89-25.96)</td>
<td>0.006</td>
<td>141</td>
<td>8</td>
<td>1.38 (0.46-4.35)</td>
<td>0.56</td>
</tr>
<tr>
<td>Hispanic Controls</td>
<td>152</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>149</td>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>White ALL</td>
<td>71</td>
<td>10</td>
<td>2.10 (0.69-7.13)</td>
<td>0.204</td>
<td>79</td>
<td>2</td>
<td>0.53 (0.09-2.94)</td>
<td>0.68</td>
</tr>
<tr>
<td>White Controls</td>
<td>82</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>83</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Asian/PI ALL</td>
<td>25</td>
<td>1</td>
<td>N/A</td>
<td>1</td>
<td>25</td>
<td>1</td>
<td>N/A</td>
<td>1</td>
</tr>
<tr>
<td>Asian/PI Controls</td>
<td>21</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>21</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Black ALL</td>
<td>10</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>10</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Black Controls</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other ALL</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Other Controls</td>
<td>3</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Logistic regression adjusted for race, age and sex  
** Logistic regression adjusted for age and sex
In utero cytomegalovirus infection and development of childhood acute lymphoblastic leukemia

Stephen Starko Francis, Amelia D. Wallace, George A. Wendt, Linlin Li, Fenyong Liu, Lee W. Riley, Scott Kogan, Kyle M. Walsh, Adam J. de Smith, Gary V. Dahl, Xiaomei Ma, Eric Delwart, Catherine Metayer and Joseph L. Wiemels