The Metagenomic Epidemiology of Childhood Leukemia

By

Stephen Starko Francis

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Epidemiology in the Graduate Division of the University of California, Berkeley

Committee in charge:
Professor Patricia A. Buffler, Chair (deceased)
Professor Steve Selvin, Chair
Professor Lee W. Riley
Professor Fenyong Liu

Spring 2014
The Metagenomic Epidemiology of Childhood Leukemia

Copyright 2014
By
Stephen Starko Francis
Abstract

The Metagenomic Epidemiology of Childhood Leukemia

By

Stephen Starko Francis

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Patricia A. Buffler, Chair (deceased)

Professor Steve Selvin, Chair

The etiology of childhood acute lymphoblastic leukemia (ALL), the most common childhood cancer, is unknown. A role for infection is supported by considerable research over the last twenty years. Despite this evidence and many infection related hypotheses, no definitive infectious agent has been identified. Until recently broad surveys of the microbiome were not possible. Modern developments in next generation sequencing and bioinformatics however now allow unbiased investigation into the astounding diversity of organisms that inhabit the human body.

This dissertation describes the first comprehensive analysis of tissue from children with childhood leukemia using metagenomic (the study of non-human nucleic acids) and epidemiologic methods. Metagenomes in pre-treatment bone marrow from the Fallon, NV ALL cluster and other ALL non-clustered cases from California were interrogated for known and novel infections. A putative and potentially novel human virus was found in the Fallon cases. To investigate viruses involved more broadly in childhood ALL, viromes of pre-treatment bone marrow from children with acute myeloid leukemia (AML) and from children with ALL (who we assumed to have similar background infections) were compared. In children with ALL, a greater intensity of viral infection as well as a significant difference in the expression of viral packets originating from double stranded DNA viruses was observed. Using 16s based bacterial metagenomics, no difference was observed in bacterial diversity between the groups showing the absence of differential contamination and suggesting that bacteria do not play a role in ALL. These results suggest a possible putative causal agent in the Fallon Nevada leukemia cluster and increased viral infection driven by double stranded DNA viruses in non-clustered children with ALL.
These data contribute to an etiologic hypothesis of ALL where a combination of pre-natal and post-natal events lead to leukemia. First, an underlying genetic predisposition, combined with an *in utero* infectious exposure, cause chromosomal anomalies and immune dysregulation. This followed by postnatal events, including toxic and infectious exposures that further modify adaptive immunity lead to a greater intensity of infections. This feeds the loop of enhanced b-cell production that facilitates virally induced self-mutation caused by foreign nucleic acids invading pre b-cells. Resulting in an ABOBEC arbitrated hypermutation. Eventually the loop of enhanced yet less competent b-cell production, RAG mediated over mutation, increased infection leading to pre b-cell viral nucleic acid loading, and ABOBEC caused self mutation remove the proliferative and apoptotic cellular checks. The result is childhood ALL.
This research is dedicated to the memory of Professor Patricia A. Buffler
She was an incredible epidemiologist, the greatest mentor and an amazing woman
I miss her
Acknowledgments

Forever thanks to my family and friends for your unwavering support and guidance through this process and my life. You are all the best I could ever wish for.

I am in debt to the California Childhood Leukemia Study for the support and resources needed to complete this work. Special thanks to Joe Wiemels, Catherine Metayer, Amelia Wallace, Jake Wendt, Theresa Saunders, Pagan Morris, Alice Kang, Todd Whitehead, Monique Does and Adam De Smith. And thanks to Blood Systems Research Institute and Eric Delwart for his guidance. Go team!

I thank my committee members Steve Selvin, Lee Riley and Fenyong Liu for their superb guidance and mentorship, and the Berkeley Epidemiology Department and the School of Public Health for providing my foundation.

I especially thank the children and families that volunteered for this study, without their participation none of this would be possible. We will not stop until we solve this mystery.
## Contents

**Chapter 1 - A new frontier**

1.1 A New Frontier ........................................... 2
1.2 Metagenomics ............................................ 2
1.3 The Human Bacterial Microbiome ................... 3
1.4 The Human Virome ....................................... 4
1.5 Microbial Balance ....................................... 5
1.6 Looking Forward ......................................... 6

**Chapter 2 - The Metagenome of Childhood Leukemia** ........................................... 7

2.1 Introduction ............................................ 9
2.2 Materials and Methods ................................. 11
   2.2.1 Fallon Viral Discovery .......................... 11
   2.2.2 ALL vs. AML Particle Isolation 454 Sequencing 13
   2.2.3 16s Bacterial Metagenomics ................... 14
2.3 Results ................................................. 15
   2.3.1 Fallon ........................................... 15
   2.3.2 Virus Particle Isolation and 454 Sequencing 16
   2.3.3 16s Bacterial Microbiome ..................... 16
2.4 Discussion ............................................. 17

**Chapter 3 - An Etiologic Framework for Acute Lymphoblastic Leukemia** ........................................... 29

3.1 On the Trail ............................................ 30
3.2 It starts from the beginning: Pre-natal Events .... 31
3.3 Pre and Post Natal Immune Development .......... 31
3.4 Post-natal Events ..................................... 32
3.5 Future Directions ..................................... 33

References .................................................. 35
Chapter 1- Metagenomic Epidemiology
1.1 A New Frontier
A new frontier of basic human biology has recently opened. Advances in next-generation nucleic acid sequencing technology are changing our understanding of what constitutes the human body and how changes in the genome could lead to disease. This new knowledge is fueling a paradigm shift from the dominant, 20th century view that viruses, bacteria, and fungi operated alone to cause disease. Now, we are beginning to view a more complex and nuanced interpretation with increased importance of commensalism, synergy, and balance of microbiota in human health (Honda & Littman, 2012). This paradigm shift affects epidemiology.

The distribution of human diseases is directly linked to how different people form groups and interact with each other and their environment. Community structure has been long recognized as a major factor in the spread of traditional infectious diseases. These diseases are very sensitive to the configuration of the social network of the population into which they are introduced (Keeling & Eames, 2005). Moreover, the importance of social networks within these networks have been documented to be important for chronic and traditionally non-infectious diseases. For example, obese individuals are more likely to have other obese individuals as contacts (Christakis & Fowler, 2007). Even in these “non-infectious” diseases microorganisms may play a greater role than previously thought. A community is comprised of groups of ecosystems any number of which can potentially be disrupted and lead to disease epidemics. An important feature of these ecosystems is its microbiome. Study of the distribution, relationship, and determinants of the microbiome structures of a community of ecosystems constitutes metagenomic epidemiology. As our ability to define the microbes in and on our bodies increases so will our ability to delineate the healthy balance of the microbiome. This will lead to an enhanced definition of the relative ‘good’ microbes from ‘bad’. With diligent study, robust comparison, and population level data, we will be able to better define those factors that influence health, disease, and community structures. Given our technical power today, the time is ripe for metagenomic epidemiology.

1.2 Metagenomics
The scale of the human microbiome is almost unimaginable. It is estimated that for every human cell there are 10 bacterial cells and for every bacterial cell there are 10 viral particles (Clemente, Ursell, Parfrey, & Knight, 2012). Such estimates have been made possible by technological spill over tools developed from human genome sequencing. Next generation sequencing technologies designed to rapidly and accurately sequence the genomes of human communities are now being used to sequence microbial communities. While these technologies continue to improve in speed and cost, the massive amount of data generated from them have produced immense challenges for our bioinformatics capabilities. Cluster based computing is essential for most metagenomic studies, although some tool kits (QIIME for example) are bringing 16s based bacterial metagenomics into the range of most researchers.
This ability to sequence all nucleic acids in a given sample without amplification of the targeted region has enabled the new field of metagenomics. Prior to these technologies culturing or PCR based amplification was required to study microorganisms. Next generation sequencing now enables unbiased identification and quantification of all known organisms in a given sample be it from the human body or the environment. As a result, a flurry of environmental microbiome studies is afoot ranging from the ocean floor to artic ice (Varin, Lovejoy, Jungblut, Vincent, & Corbeil, 2012; Yoshida, Takaki, Eitoku, Nunoura, & Takai, 2013). Metagenomics has enabled us to identify here-to-fore unknown microorganisms. In fact, this new technology has enabled the discovery of nucleic acid sequences in the human body and in environmental niches that cannot be assigned to any known microbial entities—bacteria, viruses, fungi, protozoa, or helminthes. We call these entities “dark sequences” analogous to the “dark matter” which astrophysicists hope could explain many of the mysteries of the universe. Not only are there new sequences but the quantity of these is vast. In fact what had been recognized to exist in the microbial world until 20 years ago now represents but a small microcosm of what is now known to exist today. And surely there are more to be discovered. What are they and how do they affect the human body and today's recognized microbial ecosystems remain to be determined. We have only begun to scratch the surface.

### 1.3 The Human Bacterial Microbiome:

The Human Microbiome Project (http://www.hmpdacc.org/) has contributed a strong basis for understanding ‘normal’ human flora. Recent studies show that the core microbiome is essential in metabolic and regulatory processes throughout the body (Aagaard et al., 2013). Bacteria are required for basic functions ranging from digestion of lipids and vitamins to glucose homeostasis. In addition, they can alter host gene expression, and modify drug metabolism (Clayton, Baker, Lindon, Everett, & Nicholson, 2009; Devaraj, Hemarajata, & Versalovic, 2013; Jones, Martoni, Ganopolsky, Labbe, & Prakash, 2014). Bacteria also exist as communities with structure, creating an ecosystem that synergistically interacts with the host. We are just beginning to understand the central role of bacteria in the human gut. It is now clear that our gut microbiota functions like a previously undescribed organ. Like other organs, it is essential for health and a factor in susceptibility to disease.

Most research in this arena has primarily focused on the gut bacterial flora for three reasons. First, the gut is a most rich and interactive body compartment and second, sample ascertainment is non-invasive. Finally, conducting bacterial metagenomics through 16S rRNA sequencing is relatively easy due to shared genes across bacteria and archaea. In this regard, recent technological advances (Caporaso et al., 2011) have enabled metagenomic community level descriptions for as little as $11 per sample. Such feasibility has put bacterial metagenomics into the hands of population-based epidemiologic studies.

Many of the diseases of the human intestine result from disruption of the steady-state microbial population structure with in the gut. A clear link exists between the microbiome and host immune response (Littman & Pamer, 2011). For example,
inflammatory bowel disease (IBD) has long been suspected to involve gut microbiota as some patients respond to antibiotics and fecal transplants (Brown, 2014). Indeed, IBD is associated with mutations within bacterial recognition genes (Hugot et al., 2001; Ogura et al., 2001). In this regard, obesity has been of keen interest. Animal studies have shown that transplantation of obese mouse microbiota to germ free mice can induce obesity (Turnbaugh et al., 2006), proving that certain microbiota aid in transfer of energy to the host and, intriguingly, that the microbiota are transmissible. Human twin studies have also provided interesting data. Discordantly obese mono and dizygotic twins demonstrate an association with diminished Bacteroidetes prevalence and lower overall diversity in the obese twin compared to the non-obese sibling (Turnbaugh et al., 2009). Transplanting fecal microbiota from an obese member of a human discordant twin pair to germ-free mouse caused the mouse to gain significant increase in body mass and adiposity (Ridaura et al., 2013). Further, antibiotic use during the first 6 months of life is associated with development of obesity later in life (Ajslev, Andersen, Gamborg, Sorensen, & Jess, 2011), suggesting a potentially intervenable target.

While our gut microbiota may be central players in how we develop disease, they may also be a reservoir for agents that cause disease. For example, our gut microflora act as a reservoir for the horizontal transfer of antimicrobial resistance genes (Penders, Stobberingh, Savelkoul, & Wolffs, 2013), a finding more likely to occur in food animals given low-dose antibiotics as growth promoters.

The steady-state intestinal microbial population structure is, of course, maintained by what is introduced into the intestine as food. If an ingested food product is contaminated with a human pathogen (e.g. salmonella), most of the time, the normal gut flora will protect the host against such a pathogen. However, a large inoculum of the pathogen, or a prior exposure of the intestine to an antibiotic, can overcome this protective microbial barrier. Thus, again, it is not only the pathogen per se but also the disruption of the balance in the intestinal microbiota that serves as a disease-causing factor. The association of prior exposure to antibiotics and drug-resistant salmonellosis has long been recognized (Riley et al., 1984), but now, metagenomics can provide a biologic explanation for this epidemiologic association.

Targeted study of the gut microbiome has provided astounding and novel insight into the mechanistic underpinnings of the function of the human body. In addition to the gut other body compartments (Bogdanos, Smyk, & Shoenfeld, 2013; Chen & Tsao, 2013; B. Ma, Forney, & Ravel, 2012) and other organisms are proving to harbor additional fascinating and unique properties.

1.4 The Human Virome
The totality of viruses within a human, the “virome”, is a fascinating yet challenging (complex and expensive) part of the microbiome. The evolution of viruses has likely paralleled all cellular life since the formation of nucleic acids (Koonin & Martin, 2005). Their diversity is staggering and it has been postulated that viruses are the most genetically diverse group of organisms on the planet (Suttle, 2005). This
diversity has not only brought forth some of the most pathogenic species in existence but has also made studying viral communities difficult due to extremely divergent clades of viruses that share no genetic similarity. Recent advances in metagenomics have for the first time allowed for a wide classification of all known viruses in a given sample in the greater virome. Two key findings have emerged from these studies. First we are full of both viruses of bacteria (phages) and viruses of humans, and second we are just beginning to understand the diversity of ‘normal’ human viral flora.

It is somewhat counterintuitive to accept that viruses evolve towards avirulence on an evolutionary scale. However, in maximizing fitness, pathogenicity becomes detrimental. Nevertheless, through millions of generations and unstable genomes, many viruses have “mistakenly” ventured into extremely pathogenic mutations that lead to host death. Yet pathogenicity is the exception rather than the rule. For example the Anellovirus family (or Torque Teno Virus- TTV) may in fact be the most common human virus, yet relatively little is known about them because they have not been identified as causing disease in humans. Indeed most people are likely infected with an Anellovirus, where primary infection often occurs within the first weeks of life. There is also evidence that constant transmission occurs back and forth in adults (Davidson & Shulman, 2008; TeKippe et al., 2012). The cloud of genotypes that make up TTV is highly diverse and poorly delineated. It is possible that, akin to oncogenic strains of largely benign human papilloma viruses, specific TTV strains may be associated with disease. In the instance of these highly diverse groups, traditional methods of identification are challenging. Yet metagenomics has opened the door. TTV is just one organism of many known to science that is highly prevalent yet poorly understood. Strain-specific differences in pathogenicity have been recognized, for example, between serotypes of enteroviruses or strains of papillomaviruses; some cause diarrhea or cancer and some do not. To assess the disease-causing potential of TTV and many yet-to-be known viruses will require population-based studies.

1.5 Microbial Balance

When we study humans we are, in fact, studying groups of microbial ecosystems that interact positively to promote health and negatively to promote disease. An important concept when considering the human microbiome is that of neutrality. The human body consists of groups of ecosystems that have evolved over tens of thousands of years to establish a stable equilibrium state. It is when this equilibrium state is disrupted that a disease occurs. For example, antibiotic exposure (a synthetic product) disrupts the fine balance that existed in our intestinal microbiota before antibiotics. This resulting “dysbiosis” can trigger diseases like Clostridium difficile colitis. C. diff is not a “bad” bacterium. Indeed, it is always in our intestine. It is only when other microbial populations are eliminated that C. diff becomes “bad”. Coprotransplantation works against C. diff colitis because it restores the balance. A disruption of the “local” intestinal microbiota can also affect body physiology as a whole. Other complex diseases may be caused by similar
disruptions of this fine balance – not only in the gut with bacteria but also in other body compartments and with other organisms.

The dysbiosis concept can also be extended to viruses. A primary example is that of cytomegalovirus (CMV), where prevalence in healthy adults ranges from 70-90%. Infection generally occurs early in life with minor symptoms and otherwise uninteresting clinical presentation. Yet life long latent infection ensues, a process that is the result of hundreds of thousands of years of co-evolution of humans and viruses culminating in dynamic and intricate immunologic evasion strategies. In a healthy balanced individual, CMV causes no harm, yet in an immune compromised individual, CMV can be life threatening. Dysbiosis resulting in CMV-caused morbidity can be triggered by a number of factors; some are known such as HIV and immunosuppressive therapies, yet we suspect many others are unidentified. Little is known of the viral interaction between CMV and other common viruses. Yet striking epidemiologic associations have been drawn between CMV and a broad range of complex diseases from brain tumors to heart disease (Cobbs et al., 2002; Nieto et al., 1996). For the first time we now have the ability to examine such relationships in terms of their multifactorial biological causal pathways.

1.6 Looking Forward
The microbiome may be a critical component of the epidemiology of many diseases. Genomic changes in the microbiome are measurable. It is well understood how pathogens are transmitted between individuals within communities. Through our growing understanding of the metagenome, it is clear that we not only transmit pathogens, but we also transmit our commensal flora as well (Unpublished data). Through direct or indirect contact, we may modify disease risk. In addition to the dysbiosis described above, additional modification may occur through more traditional exposures:

**Chemicals**- Immunotoxic mechanisms alter host immune response and therefore microbial communities.

**Stress**- A psychologically-induced response that can be immunosuppressive and alter our core microbiome or risk of dysregulation.

**Contact**- Direct or indirect contact modification of microbiota is occurring on a daily basis.

**Timing and frequency of infection** - Early life exposures to both commensal and pathogenic organisms mold our immune system for the remainder of the life course. **And others....**

The combined tools of metagenomics and epidemiology will lend insight into human health and disease. As laboratory and bioinformatics methods become streamlined, defining and quantifying known organisms, as well as the expanse of unknown organisms and how they interact, will likely become a critical aspect of this field. Challenges including intra-subject variation and temporal stability will require multiple sampling schemes. Despite these challenges, paired with detailed genetic predisposition, environmental and social exposures, metagenomic
epidemiology may very well be the key to untangling decades of cryptic disease associations. We are excited by this new tool in our epidemiologic quiver.
Chapter 2- The Metagenome of Childhood Leukemia
2.1 Introduction:
The role of infection in the etiology of acute lymphoblastic leukemia (ALL) in childhood has been the focus of hypotheses and inquiries for decades. This study represents the first comprehensive investigation of possible etiologic viruses and bacteria.

Acute lymphoblastic leukemia (ALL) is the most common subtype of the most common childhood cancer. Within common ALL (cALL), a pre B-cell ALL occurring between the ages of 2 and 5, the presence of MLL-AF4 and TEL-AML1 fusions present in blood spots taken at birth from children that go on to develop this common type of leukemia provides unambiguous molecular evidence for initiation in utero, (Maia et al., 2003; Wiemels et al., 2002). However these translocations are ~100 times more common than disease making them necessary, but not sufficient to cause disease(M. Greaves, 2006a). Furthermore, recent studies have shown a profound IL-10 deficit in children that later develop leukemia, providing evidence of immune modification at birth and further supporting in-utero initiation (Chang et al., 2011). Together, these observations necessitate a post-natal ‘second hit’ for development of disease.

A strong body of evidence suggests a role for infection in the etiology of ALL, specifically as a disease initiating ‘second hit’. Two main infection hypotheses have been set forth, Greaves’ delayed infection model (M. Greaves, 2006b) and Kinlen’s population mixing model (Kinlen, 1998). While both of these hypotheses explain different aspects of epidemiologic observations, the models are not mutually exclusive. Greaves’ model proposes that delayed exposure to infectious agents leads to an abnormal immunologic response promoting the secondary genetic lesions that lead to disease. Kinlen’s model proposes that leukemia is the result of a rare response to common infectious agents. The observations upon which these hypotheses were built have since been greatly expanded. Most notably, it is now clear that children that develop leukemia have more medically diagnosed infections prior to the diagnosis of leukemia than healthy children (Chang, Tsai, Tsai, & Wiemels, 2012; Crouch et al., 2012). It is unclear if these children are infected at a greater frequency or if they have more severe infections warranting a visit to a physician. The observation of decreased IL-10 levels at birth suggests that children who develop leukemia are born with immunologic defects that predispose them to greater and/or more severe infections throughout their life course prior to diagnosis. Yet a central question is, what are these infections? Is the infection a transforming virus, or a suite of infections that modify the immune system and initiate a cascade of events that establish a pre-leukemic clone? Both questions remain of keen interest.

Both PCR and immunologic assays have been used to investigate specific pathogens including bovine leukemia virus (Bender et al., 1988), polyoma viruses (MacKenzie,
Perry, Ford, Jarrett, & Greaves, 1999), parvo B19 (Heegaard, Jensen, Hornsleth, & Schmiegelow, 1999; Heegaard & Schmiegelow, 2002; Isa, Priftakis, Bröladen, & Gustafsson, 2004) and human herpes virus (MacKenzie et al., 2001). The most comprehensive study to date utilized a representative difference analysis (RDA) but found no foreign detectable sequences in 11 cALL subjects (MacKenzie et al., 2006). Yet hybridization-based RDA has many potential pitfalls not present in modern deep sequencing based analyses. While no previous study has identified a single causative agent, all fail to prove the contrary. Investigating multiple infections adds additional challenges. Combining epidemiologic techniques with laboratory methodologies similar to those presented in this paper provide a substantial leap in current knowledge regarding this complex mechanism.

Both Kinlen and Greaves’ hypotheses also seek to explain the occurrence of clusters of childhood leukemia. The significance of leukemia clusters remains a controversial subject (Goodman, Naiman, Goodman, & LaKind, 2012). While many clusters have been defined, the most poignant example in the United States occurred in Fallon, NV. From 1997-2001, 17 children were diagnosed with childhood leukemia and 9 of these cases occurred in a narrow 11-month window (Francis, Selvin, Yang, Buffler, & Wiemels, 2012). We have previously conducted a spatial analysis of the cluster cases and found a non random distribution of cases around Fallon, rather than in the urban center. In other words, cases occurred primarily in rural areas where land is used for agricultural purposes. One such use is flood-irrigated farm land. Fallon has >50,000 hectares of such land, an ideal habitat for breeding mosquitoes. Given an extremely high infection rate in this area with other vector-borne pathogens, such as West Nile Virus, we hypothesize that the Fallon cluster may have been the result of an unidentified virus, transmitted by a mosquito vector, causing the so-called ‘second hit’ event for children living around Fallon, who were already at risk for leukemia. If a causal virus is implicated in the Fallon cluster, it would likely be unique to that cluster, or to leukemia clusters in general and not a universal etiologic agent of ALL. Therefore we also investigated the more prevalent sporadic cases of ALL.

To investigate the ALL infection hypotheses we conducted a 4-part study comprising the most comprehensive viral discovery and bacterial quantification project in ALL to date. In addition to comparing clustered to non-clustered subjects, we compared non-clustered ALL cases to acute myeloid leukemia’s (AML) drawn from the same underlying population at the same time. We selected AML subjects as a comparison group because AML as a cancer has vastly different epidemiologic patterns, no shared genetic risk alleles, and an early differentiation in lineage from hematopoietic stem cells yet provides and estimate of infection exposure of another type of immune suppressed child to reduce the background infection ‘noise’. We also sought to identify the infection burden in bone marrow of children with ALL.

1- In-depth virome analysis of pretreatment bone marrow from 2 subjects from Fallon and 2 non-clustered ALL subjects

10
2- Comparison of all viral particles in pretreatment bone marrow of 36 ALL and 11 AML subjects using particle isolation and 454 sequencing
3- Comparison of 16s bacterial metagenomics pretreatment bone marrow of 73 ALL and 28 AML subjects.

2.2 Materials and Methods:
The University of California Institutional Review Board and all participating institutions approved this study. Informed consent was obtained from all participating subjects.

2.2.1: Fallon Viral Discovery
Sample Selection:
Two pretreatment bone marrow samples were obtained from the Children's Oncology Group (COG). The eligibility criteria and DNA collection procedures have been previously described (O'Leary, Krailo, Anderson, Reaman, & Children's Oncology, 2008). These were 2 of 3 samples collected during routine COG collection as part of the Fallon Nevada leukemia cluster Center for Disease Control investigation. These samples met eligibility criteria for admission into COG and were not collected specifically as cluster samples, yet were involved in the cluster.

Two additional comparison pretreatment bone marrow specimens were selected from the California Childhood Leukemia Study (CCLS). The CCLS has been described elsewhere (X. Ma, Buffler, Layefsky, Does, & Reynolds, 2004). Briefly, the CCLS is an ongoing case-control study in California that commenced in 1995. The study began in the 17 counties surrounding the San Francisco Bay Area, and now encompasses most of California through collaboration with 35 pediatric oncology centers in the state, allowing for rapid case ascertainment and biospecimen retrieval, usually within 72 hours from diagnosis. One or two controls are recruited using birth certificate information from the California Department of Public Health Office of Vital Records. Controls are matched on age, sex, Hispanic ethnicity, and maternal race.

These pretreatment samples were matched on month of diagnosis to the two Fallon samples (1999-2000). These samples were not part of any identified cluster within the CCLS.

RNA extraction:
All laboratory equipment and reagents were thoroughly cleaned and decontaminated to minimize contamination potential. Ultra clean reagents along with liberal use of UV and chemical treatment (Bleach solution and RNase away) were used to ensure that no contaminating sequences entered pre-extracted or post-extracted samples.

The 4 bone marrow specimens were extracted using the same protocol in the same decontaminated hood at the same time. We used a modified protocol for trizol RNA extraction with back extraction for DNA. A 250ul aliquot of bone marrow was
suspended in 750uL Trizol LS. After incubating for 5 minutes at room temperature, 200uL chloroform was added and shaken vigorously by hand for 15 seconds. After incubating for 2-15 minutes at room temperature the sample was centrifuged at 12,000 xg RCF at 4°c for 15 minutes. The aqueous phase was then removed and 0.5uL glycogen was added and mixed with pipette. 500uL of 100% isopropanol was added and incubated at room temperature for 10 minutes. The sample was then centrifuged at 12,000 xg RCF at 4°c for 10 minutes. The supernatant was removed and the RNA pellet was washed with 1.0mL 75% ethanol then vortexed and centrifuged at 7,500 xg RCF, 4°c for 5 minutes. The supernatant was removed and RNA pellet dried using vacuum. The RNA pellet was then re-suspend 30uL RNase free water. RNA quantity and quality was analyzed using Agilent Bioanalyzer.

**Library preparation and ribosomal capture:**

Prior to Illumina library preparation, a ribosomal capture was performed to reduce the amount of human ribosomal transcripts being sequenced. This was completed to increase the sensitivity of identifying non-human transcripts through physical removal of ribosomal sequences prior to sequencing. For each RNA sample, ribosomal RNAs were depleted using Ribo-Zero™ Gold Kits (Epicenter) & the depleted RNA was confirmed with Bioanalyzer. RNA-seq libraries were prepared using PrepX RNA-Seq Library kits (WaferGen, Inc) according to the manufacturer's protocol. The library was PCR amplified & the reaction was cleaned with AGENCOURT® AMPURE® XP (Beckman).

**Sequencing/Quality Filtering:**

Sequencing was preformed on an Illumina Hiseq2000. Each sample generated approximately 60 million high quality 100bp paired end reads per sample. Those reads were aggressively filtered using both Illumina’s CASAVA 1.8 ‘pass/no pass’ system then were further quality filtered using the fastx tool kit to only include reads with a minimum phred score of 30 over at least 75% of the read.

**Bioinformatics:**

A custom bioinformatics pipeline was developed for identification of known and unknown non-human nucleic acids. The system, named DARK, preforms computational human subtraction, contiguous
sequence (contig) assembly and sequence identification designed for use on SLURM (https://computing.llnl.gov/linux/slurm/) based computer clusters. Briefly after aggressive quality filtering DARK iteratively aligns the raw 100bp reads using BowTie2 (Langmead & Salzberg, 2012) to 3 human reference genomes and the Refseq human database. All human reads are removed. DARK assembles the raw non-human 100bp reads using Trinity (Haas et al., 2013) into longer contigs. Both the raw non-human 100bp reads and contigs are separately queried against Genebank (https://www.ncbi.nlm.nih.gov/genbank/) and the refseq virus database (https://www.ncbi.nlm.nih.gov/genbank/) using BLAST (Altschul et al., 1997). This is accomplished through massively parallel computing on an in-house SLURM based cluster. To increase our specificity at the cost of sensitivity in the current analyses, we chose to focus on contigs assembled by Trinity. By using contigs instead of raw 100bp sequences, we are reducing the probability of sequencing error resulting in false hit by requiring support for every base pair and by increasing the size of sequences queried in blast. Although this may reduce our sensitivity, this is a conservative approach.

Analysis of Blast files:
We used a combination of approaches to examine the resulting BLAST files produced by our DARK pipeline. We designed a custom database to parse and query blast files by subject and read. This enabled filtering of hits according to a high discriminatory likelihood and collating hits by subject. We utilized MEGAN, a freely available package that complies visualization of metagenomic data (Huson, Auch, Qi, & Schuster, 2007).

2.2.2 ALL vs. AML Particle Isolation 454 Sequencing
Sample Selection:
All samples were obtained from the CCLS (described above). We selected 36 cALLs (pre-B cell ALL with CD 10:19, age at diagnosis 2-5.9 years) and 11 AML’s with a similar age distribution. All samples were drawn from the same underlying study population occurring in the 17 counties surrounding the San Francisco bay area. Pretreatment bone marrow specimens were obtained from the original diagnostic bone marrow aspirate. The existing CCLS study allowed for rapid case ascertainment of both biospecimens and epidemiologic data.

Viral particle purification:
Bone marrow specimens were suspended in Hanks’ balanced salt solution (Gibco BRL). Three 15,000 g centrifugations using a tabletop microfuge and a0.45-um filter (Millipore) to remove non-virus sized particles. The filtrate was treated with a mixture of DNases (Turbo DNase from Ambion, Baseline-ZERO from Epicen-tre, and Benzonase from Novagen) and RNase (Fermentas) to digest all nucleic acids not protected by viral capsids. All remaining protected viral nucleic acids were then extracted using the QIAamp viral RNA and DNA extraction kits (Qiagen). Viral cDNA synthesis from extracted viral RNA/DNA was performed as described previously (Victoria, Kapoor, Dupuis, Schnurr, & Delwart, 2008).
Bioinformatics:
After sequencing, raw data generated from a Roche 454 sequencer data was quality filtered for sequencing errors and haplopolymer regions then converted to fasta files using the fastx toolkit. The resulting fasta files were parsed and trimmed by subject based on the unique barcode ligated to each read.Parsed fasta files were queries using BLAST against both the NT and viral refseq databases using BlastN and BlastX. Recovered nucleic acids that were assigned a virus or virus family based on E-values tabulated and compared between groups. The E-Value cutoff for calling a hit to a virus was set at $E^{-10}$, a highly stringent cutoff. The number of viral hits was normalized by subject based on number of sequence reads recovered from each sample, the number of viral hits was simply divided by the total recovered sequences for a given sample. Wilcoxon rank sum tests were used to compare mean frequencies of called viral hits between ALL and AML subjects. These were graphically displayed using 95% confidence intervals. All statistical analyses were preformed in R.

2.2.3 16s Bacterial Metagenomics

Sample selection:
All samples were drawn from the CCLS study base and biorepository as described above. We analyzed bone marrow from 73 cALLs and 28 AMLs.

DNA Extraction:
DNA was extracted using a modified TRIzol LS protocol. During initial RNA separation described above any remaining aqueous phase was removed, 375uL back extraction buffer (4 M guanidine thiocyanate, 50 mM sodium citrate and 1 M Tris pH 8.0) was added, mixed and incubated for 10 minutes at room temperature. The samples were then centrifuged at 12,000 xg RCF at 4°C for 15 minutes. The aqueous phase was removed and combined with 0.5ul glycogen to and mixed with a pipette. DNA was precipitated with -260uL 100% isopropanol (~80% volume of aqueous phase) and incubate for 5 minutes at room temperature, and mixed occasionally. Then the samples were centrifuge at 12,000g RCF at 4°C for 5 minutes, the supernatant was removed then pellet washed with 1.0mL 75% ethanol and incubated at room temperature for 15 minutes while mixed occasionally. Samples were then centrifuged at 2,000g RCF at 4°C for 5 minutes, the supernatant was removed and DNA pellet was dried. The dry pellet was re-suspend in 30uL of nuclease-free water.

Bacterial 16s PCR and sequencing
To examine the diversity and prevalence of bacterial infection in bone marrow we performed PCR amplification using barcoded primers specific to the 16s rRNA gene shared by all bacteria, we modified the Caporaso et al. protocol(Caporaso et al., 2011). Briefly, custom barcoded primers were constructed targeting the V4 region of the 16s gene. A unique 12-nucleotide barcode was constructed by synthesizing additional nucleotides to existing V4 primers. One hundred unique barcoded
primers were constructed. Barcodes were optimized for error reduction (Caporaso et al., 2010). We did not include the Illumina sequencing adapter and linker on our primers. Each of these unique primers was used to amplify the V4 region of 16s in each sample, creating a uniquely barcoded amplicon. The PCR was performed in triplicate.

**Sample pooling, library preparation and sequencing:**
Samples were normalized and pooled using pico green in triplicate. All samples were diluted to the lowest concentration sample. The pool was then prepared for sequencing by the Berkeley Functional Genomics core facility, where an Illumina sequencing pad, linker and adapter was ligated to each amplicon. Sequencing was conducted at the Berkeley QB3 sequencing core facility. An Illumina HiSeq2500 was used to produce 100bp paired end data.

**16s bioinformatics:**
Sequence pre-processing and OTU analysis was performed using the FastX Toolkit ([http://hannonlab.cshl.edu/fastx_toolkit](http://hannonlab.cshl.edu/fastx_toolkit)) and QIIME (Caporaso et al., 2010). Quality filtering was performed where a minimum 30 quality score over at least 75% of the sequence read, no ambiguous bases, and 1 primer mismatch was allowed. Quality sequences were then parsed by barcode and binned into OTU’s. A subsample from each subject of 20,000 sequences was used for OTU construction at a 97% sequence similarity cutoff using uclust. Taxonomic classification was performed by querying the Greengenes database (2011 release) (DeSantis et al., 2006).

**2.3 Results:**

**2.3.1 Fallon:**
Approximately 299 million 100bp paired end reads were recovered that passed our quality filters. DARK then removed 217 million reads as human during the first rounds of BowTie2 based computational subtraction, leaving 81 million reads. Those raw reads were denovo assembled into 165,072 contigs using Trinity. Using BLAST we were able to further classify 136,916 as human or low complexity. Of those non-human contigs we were able to categorize 4.2% as virus, 23.4% as bacterial using both Blastn (Figure 1). The remaining non-human unknown contigs, the dark sequences, made up the remaining 72.5% and are of keen interest yet not explored in this paper.

Of the identified non-human sequences, both the Fallon subjects and the CCLS subjects share many of the same viral infections (Figure 2). Each sample contained phages belonging to the Caudoviridae and Baculoviridae families and are likely the result of bacterial infection or contamination from the initial lumbar puncture to obtain bone marrow aspirate (see 16s section). There are many reads that hit to double stranded DNA viruses, specifically herpes viruses. All 4 samples contained strong evidence of Cytomegalovirus (CMV) infection. Human endogenous retrovirus (HERV) sequences were also recovered from all 4 subjects, with the most prevalent HERV species identified being a member of the HERV K family.
The only group of viruses found strictly in the Fallon subjects and not the sporadic CCLS cases are members of the ssRNA virus family. The contigs had high homology bovine viral diarrheal virus (BVDV), a Flavivirus and common pathogen in cattle. Although these reads are strong hits to BVDV there is evidence in the raw files that other ssRNA viruses may be present.

2.3.2 Virus Particle Isolation and 454 Sequencing
After sequencing, quality filtering and parsing, ~300-4000 reads were recovered per subject. BlastN and tBlastX results noted a number of viral hits in both ALL and AML subjects and significant differences were observed using Wilcox ranked sum test between the groups (Figures 3 & 4). The most prevalent viruses were members of the double stranded (DS) DNA virus family, specifically herpes viruses. All subjects had evidence of DS DNA virus infection. A statically significant difference was observed for DS DNA viruses between ALL and AML subjects (BlastN p-value 0.0007, tBlastX p-value= 0.023). HERVs were also detected in both groups, and a marginally significant difference was noted between ALLs and AMLs (BlastN p-value=0.058, tBlastX p-value=0.143). There was also evidence of retrovirus infection. Yet the number of recovered reads was low and misclassification of HERV sequences is likely and there was no observed difference between groups (BlastN p-value=0.089). Both DS RNA and single stranded (SS) RNA viruses were detected in low levels in both AML and ALL subjects though there was no significant difference between the groups (DS-RNA BlastN p-value= 0.09, DS-RNA tBlastX p-value= 0.158, SS-RNA BlastN p-value= 0.65, SS-RNA tBlastX p-value= 0.181). Overall, when examining mean load of all identified viral sequences meeting the E-value cutoff we observed a significant difference between ALL and AML subjects (BlastN p-value= 0.013, tBlastX p-value= 0.046).

2.3.3 16s Bacterial Microbiome
After sequencing, 28,775,627 reads passed initial Illumina CASAVA 1.8 quality filtering. After barcode parsing and rarefaction we utilized 20,000 reads per subject. Those sub sampled reads were assigned OTUs in the QIIME pipeline. Figure 5 shows the bacterial composition at the class level for each group, where Figure 6 shows individual level data. The most common bacterial class was Gammaproteobacteria (ALL=51.70%, AML=49.37%), followed by Betaproteobacteria (ALL=23.34%, AML=25.72) and Flavobacteria (ALL=9.24%, AML=10.98) then Alphaproteobacteria (ALL=5.18%, AML=4.27%) being the last class above 5% of the sample. Additional classes of bacteria under 5% are detailed in Figures 5-7. When examining bacterial classifications at the most granular level possible without misclassification, by genus, the most common bacteria was Pseudomonadaceae (ALL=45.31%, AML=42.56%), followed by Comamonadaceae (ALL=19.51, AML=21.06%), Weeksellaceae (ALL=9.23%, AML=10.96%), and Moraxellaceae (ALL= 3.61%, 4.35%). The many other genus's representing groups less than 4% (detailed in supplemental material).
No evidence of differences between quantity of bacterial classes or genus’s using Wilcoxon sign-rank test while adjusting for multiple comparisons was observed. There is no evidence for a single bacterial genera present differentially between ALL’s and AMLs (Figure 6 & 7). Furthermore, when examining neighbor joining trees produced by of the OTU’s assigned with in each samples, there is no evidence for any difference between ALL’s and AML's (Figure 8).

2.4 Discussion:
This is the first investigation to comprehensively describe the non-human nucleic acids present in pretreatment bone marrow specimens of childhood leukemia using new nucleic acid sequencing tools. This study was designed to investigate the two main etiologic hypotheses postulated by Kinlen and Greaves that stipulate the role of infection in the etiology of ALL. We examined metagenomes in the bone marrow of subjects from a defined cluster of ALL (Fallon), non-clustered ALL subjects (CCLS), and other immnosuppressed children (AMLs).

This metagenomic investigation demonstrates that children who have either lymphoid or myeloid leukemia harbor diverse viral and bacterial communities with in their bone marrow. This observation should come as no surprise as bone marrow from healthy individuals harbor diverse bacteria and viruses. Nevertheless, our comparisons show clear differences in the microbiome of subjects with lymphoid and myeloid leukemia. There are 3 main findings from this investigation:

1- Subjects from the Fallon leukemia cluster were infected with pestiviruses not found in the non-clustered subjects examined.
2- ALL subjects have a greater infection load, specifically of Herpesviruses than AML subjects
3- No detectable differences in bacterial communities between ALL and AML subjects were detected.

The Fallon leukemia cluster is likely the most significant clustering event of leukemia known to the medical community. Leading infection hypotheses state that clusters of leukemia may be the result of a transient virus spreading through a community, causing leukemia in susceptible children. If that were the case, we would expect to detect the causal virus in cluster cases and not in non-clustered, non-related leukemia subjects. It is that prediction and our subsequent observation of a ssRNA virus, definitely a pestivirus likely a flavivirus, in the Fallon subjects and not any other ALL tested that deserve further consideration.

A flavivirus or other vector borne virus may play a role in ALL. The flavivirus, related to BVDV, causes birth defects in cattle(Lindberg, 2003). Related viruses show the ability to induce genomic instability. Also, flaviviruses can be spread by vector means. In line with a previous hypothesis set forth by our group wherein we suggested a possible vector route of infection in the Fallon leukemia cluster(Francis et al., 2012). Taken together, previous study highlighting the abrupt epidemic curve of the cluster and the geographic concordance with mosquito breeding habitat and captured mosquitoes along with, this new observation suggesting the presence of a virus with vector borne potential in the bone marrow of clustered ALL subjects add
Evidence to the concept of a vector-borne virus in the causal chain of some children with ALL. Though, our ability to only obtain two viable samples limits inference greatly. Further studies of additional Fallon samples must be conducted.

If indeed the Fallon cluster was the result of vector-borne agent and that agent was unique to that cluster, or clustered cases generally, what may be the role of infection for the majority of ALLs, non-clustered ALLs? We suspect that clusters are the result of a novel infection transiently elevating leukemia risk, where other potentially common infections may set in motion the immunologic damage that generates the leukemic clone. Common infections may pose differential leukemia risk by timing of infection as set forth by Greaves (M. F. Greaves & Alexander, 1993). We chose AML subjects as a comparison group in order to estimate infection exposure of another type of immune suppressed child to reduce the background infection ‘noise’. What we observed was a striking difference between ALLs and AMLs in terms of infection load with any virus, and more specifically a statistically significant difference in the amount of herpesvirus detected in pretreatment samples. These results are similar to a previous more narrow herpes specific study showing a greater infection burden of CMV in B-cell leukemia’s (Hermouet et al., 2003).

Herpes viruses are complex viruses with highly evolved large genomes that harbor many immune evasion genes, most of which are poorly understood (Nicholas, 2000; Vossen, Westerhout, Soderberg-Naucler, & Wiertz, 2002). Some viruses in the herpeviridae are nearly ubiquitous while others remain relatively rare. Herpes viruses such as Kaposi sarcoma virus, Epstein-Barr virus, and cytomegalovirus have oncogenic potential, additionally their large repertoire of immune evasion genes suggests that immune disregulation is a critical aspect of herpes virus life cycles (Nicholas, 2000). Herpesviruses share many genes across species, and many repeat regions exist within their genomes (Davison, 2007). These repetitive regions are thought to be involved non-redundantly during circularization, yet their function is largely unknown (Davison, 2007; Nicoll, Proenca, & Efstathiou, 2012). Related to these repeats, determining the exact herpes virus where our recovered sequences originated is challenging. Despite these challenges it is clear that many of the recovered sequences originate from cytomegalovirus (CMV). Human herpesvirus-5 or CMV infects 50-80% of adults in the United States and >90% worldwide. While it is a common infection with typically mild symptoms complications can occur with immune suppression. CMV is highly evolved to remain latent, yet if that latency is modified by immune suppression it has the ability to cause serious harm to the host. Therefore whether the measured difference in herpesvirus between AMLs and ALLs is related to cause or is simply the result of differential immune suppression between the groups is a valid question needing further investigation. Yet, evidence exists that a difference in herpesvirus infection may exist before disease initiation. Maternal infection during pregnancy, as measured in 342 case mothers and 1,216 control mothers as measured by serum EBV IgM at 14 weeks has been associated (OR = 2.9, p=0.0002) with childhood ALL (Lehtinen et al., 2003). Additional studies of the timing and immune response to herpes virus infection and its role in childhood leukemia is warranted.
Prior evidence of the existence of oncogenic viruses make a viral agent the most plausible candidate for an infectious etiology of ALL, yet since the discovery of a causal role of Helicobacter pylori and stomach cancer the role of bacteria as an etiologic agent in cancer has gained traction. To examine if a candidate bacterial agent could be identified in ALL we used a similar theory to the viral discovery comparison, utilizing AMLs as an immune suppressed control to estimate the infection exposure history of an immune suppressed child. After performing barcoded v4 16s PCR and subsequent analysis we found no difference between the ALL and AML groups. While this similarity suggests that there is no etiologic role for bacteria in leukemia, there may be clinical relevance to the description of the type of bacteria found in the pretreatment bone marrow. Yet, inference is difficult. Although the lumbar area was sterilized prior to the bone marrow aspirate sampling any bacterial DNA would still be present. We would therefore be unable to distinguish between skin contamination from the lumbar puncture and true bone marrow infection.

Despite this limitation the findings from this 16s study are valuable for drawing inference from the viral discovery portion of this investigation. During the analysis phase of the 454 viral discovery data, concern was raised since herpes viruses share genes and repeat regions with a variety of bacteria and phages. We now have evidence that there are equal bacterial prevalence and diversity between ALL and AML leukemia’s suggesting that differential contamination has not occurred and that our measured viral differences between ALLs and AMLs are likely not the result of bacterial genes recovered from phages in the particle isolation viral discovery.

There are significant limitations to this study. The most in-depth viral discovery work was conducted on 2 Fallon subjects and 2 CCLS subjects. These subjects were studies with the greatest sequencing effort and therefore are subject to the greatest scrutiny. Although, different viral sequences were present in the Fallon subjects it is difficult to draw inference due to the small sample. Indeed, with randomly drawn subjects it would not be unlikely to obtain different infection patterns that may converge by chance. Yet in the analysis of the contiguous sequences it is clear that the Fallon ALLs share most of the same viruses with the CCLS ALLs, only the flavivirus group is strictly in the Fallon ALLs. The particle isolation and 454 sequencing of ALLs and AMLs also has limitations. The statistical power was reduced by the small sample, thus the significant results obtained are a testament to the large observed difference between groups. The lack of a healthy control group also limits inference, although a proper control in this case is a difficult question. Obtaining bone marrow from healthy children is ethically challenging, yet the primary reason for using AMLs as a comparison is that we believe AMLs better represent the infection exposure history of an immune suppressed child. If we were to compare the virome of ALL cases to healthy controls we would expect to find many more infections present in the ALL’s simply due to the result of an ineffective immune system at the time the samples were taken. By using AMLs as controls we are reducing that signal-to-noise ratio by controlling for infections common to immune suppressed children. In effect, we hope that this comparison will improve
our inference by helping with inherent issues of temporality when searching for infections in diagnostic samples. Regardless, having a healthy control population would greatly improve the generalizability of this study. Finally the bacterial analysis highlights some of the sample collection difficulties. With new and highly sensitive metagenomic techniques, methods of sample collection that eliminate contamination from different body sites are challenging. In this case, the lumbar puncture to obtain bone marrow undoubtedly introduced skin microflora into the bone marrow sample making the skin microbiota and the bone marrow microbiota indistinguishable. We are unable to say with certainty that any of the bacteria measured originated in bone marrow. Yet the observed similarity between AML and ALL subjects strongly suggests that contamination was equal between groups.

In conclusion we have completed a first look into the metagenome of children with leukemia. Our results suggest that children from the Fallon leukemia cluster may harbor a virus that other children with leukemia are not infected with. That virus is likely a member of the flavivirus family and may have been transmitted by mosquito. Further laboratory studies are needed to determine the exact virus and general prevalence in leukemia subjects and healthy individuals. Children with ALL have a different virome than children with AML. Children with ALL have more viral nucleic acids protected by capsids in bone marrow than children with AML and most of that difference is the result of herpesviruses. There is no evidence of a difference in the bacterial communities of bone marrow specimens between ALL and AML subjects. These results are far from conclusive yet provide the window into the metagenome of childhood leukemia. Clearly, we have generated more questions than answers and detailed follow up is needed. Is the ssRNA virus unique to Fallon? Do healthy individuals around Fallon harbor the same virus? Do ALL subjects express more herpes viruses than AMLs in a replication set? Which herpes virus specifically? When are ALLs infected with herpes and is the timing and prevalence of infection different than healthy children and AMLs?

To answer these questions additional metagenomic studies should be conducted in childhood leukemia and healthy subjects to define the timing, prevalence and progression of infection from conception into adulthood. These mechanisms may indeed be fundamental to not only understanding ALL, but also understanding both health and disease.
Figure 1: BLASTN Results of High Complexity Non-Human Contiguous Sequences
Figure 2: Comparison of the Virome Between Fallon and CCLS Subjects. BLAST results from Trinity assembled contiguous sequences. Red = CCLS Subject 1, Blue = CCLS Subject 2, Green = Fallon Subject 1, Yellow = Fallon Subject 2.
Figure 3: BLASTn Viral Metagenome Comparison ALL vs. AML
BLASTn (E-Value Cutoff= $E^{-10}$) classification of viral capsids sequenced.
Figure 4: tBLASTx Viral Metagenome Comparison ALL vs. AML
tBLASTx (E-Value Cutoff= E^{-10}) classification of viral capsids sequenced.
Figure 5: Bacterial 16s Classification (Class) of Grouped ALL's (73) and AML's (28)
Figure 6: Bacterial 16s Classification (Class) of Individual ALL’s and AML’s
For legend see Figure 5
Figure 7: Heatmap of Bacterial 16s Classification (Class) of Individual ALL's and AML's
Figure 8: Neighbor Joining Tree of 16s Defined Genera
Chapter 3 - An Etiologic Framework for Acute Lymphoblastic Leukemia
3.1 On the Trail

Despite decades of major advances in treatment, acute lymphoblastic leukemia (ALL) remains the most common childhood cancer. Although clinical resolution can be obtained, the treatment itself and the long-term effects are substantial (Hudson et al., 2013; Oeffinger et al., 2006). Unfortunately the underlying cause of ALL remains largely a mystery. We present here an etiologic framework for ALL, a combination of evidence and hypothesis that we hope will guide research toward the understanding of the cause of ALL. Furthermore we hope that this framework will identify research focused on etiologic mechanisms that are viable intervention targets with the ultimate goal of preventing the disease.

Hypothesis:

Here it is hypothesized that a combination of pre-natal and post-natal events lead to ALL (see figure). First, an underlying genetic predisposition, combined with an *in utero* infectious exposure, cause chromosomal anomalies and immune dysregulation. This followed by postnatal events, including toxic and infectious exposures that further modify adaptive immunity leading to a greater intensity of infections. This feeds the loop of enhanced b-cell production that facilitates virally induced self-mutation caused by foreign nucleic acids invading pre b-cells. This results in an ABOBEC arbitrated hypermutation. Eventually the loop of enhanced yet less competent b-cell production, RAG mediated over mutation, increased infection leading to pre b-cell viral nucleic acid loading, and APOBEC caused self mutation remove the proliferative and apoptotic cellular checks. The result is childhood ALL.
3.2 It starts from the beginning: Pre-natal Events
Underlying heritable polymorphisms in *IKZF1, CEBPE* and *ARID5B* have been identified in childhood ALL (Trevino et al., 2009). Yet a low concordance, as well as clear evidence of a single clonal origin in concordant cases, between monozygotic twins nearly ensures a non-heritable exogenous trigger (M. F. Greaves & Alexander, 1993). Some subtypes of ALL, including the most common subtype (~24%) with an *TEL-AML1* (or *ETV6-RUNX-1*) translocation, are initiated in utero as demonstrated by the presence of leukemic B-cell chromosomal lesions in new born blood spots (NBS) collected directly after birth (Wiemels et al., 1999). Studies have shown that these specific lesions at birth, are ~100 times more prevalent than leukemia itself, therefore, these chromosomal lesions are likely a necessary but not sufficient cause of ALL (Mori et al., 2002). This observation laid the foundation for Greaves model (M. Greaves, 2006a; M. F. Greaves, 1988) which proposed a secondary postnatal 'hit' triggering disease in susceptible children.

Further, recent studies have shown a deficit of the anti-inflammatory cytokine IL-10 in NBS at birth in children that go on to develop leukemia compared to healthy controls suggesting that immune dysregulation is present at birth (Chang et al., 2011). Whether this deficit is due to other genetic lesions, maternal /fetal immune receptor interactions (such as KIR-HLA) or exogenous exposures (such as in utero infection) is unknown.

3.3 Pre and Post Nataal Immune Development
This underlying damage to adaptive immunity may result in errors in antigen recognition. For instance the Tel-AML1 fusion results in slowing of b-cell differentiation. During maturation b-cells are detained in a critical stage, RAG mediated VDJ recombination, creating additional recombination events. During B-cell development an extraordinary stochastic shuffling of specific immunoglobulin (*IGH*) or T cell receptor (*TCR*) loci occurs, the purpose being expansion of the 5immune repertoire by randomly creating new *antigen receptor* loci that could recognize novel immunologic challenges. This targeted recombination by RAG proteins is highly regulated because of the inherent danger of a mutation driven process, yet the TEL-AML1 fusion appears to allow for enhanced mutation It is estimated that $1 \times 10^{11}$ B-cells are created every day, yet normally only $1 \times 10^9$ survive the process, this may be drastically different in TEL-AML. These errors result in B-cells that would have sterile antigen recognition and fail to undergo clonal expansion during times of response to infection. This would likely facilitate additional B-cell production therefore producing a greater opportunity for RAG mediated VDJ recombination. An increase in B-cell production and RAG activity would result in a greater chance of deleterious RAG mediated mutations.
3.4 Post-natal Events

Postnatal events associated with ALL have been identified. One example of a possible postnatal modifying event comes from a recent paper suggesting that Cesarean delivery increases risk of childhood ALL; possibly by virtue of differential primary population of the human microbiome compared to children born by vaginal delivery. This change may lead to postnatal endogenous immune modification that would increase susceptibility, Indeed it is becoming widely accepted that correct priming of the immune system early in life is critical for proper immune development.

Toxic exposure may also facilitate dysregulation of adaptive immune response in ALL. This far-from-novel concept is supported by decades of research detailing the ability of immunotoxic compounds to alter immune function (Corsini et al., 2011; Galloway & Handy, 2003; Gleichmann, Kimber, & Purchase, 1989; Holladay & Smialowicz, 2000; Voccia, Blakley, Brousseau, & Fournier, 1999; Vojdani, Ghoneum, & Brautbar, 1992) through a variety of mechanisms. Extreme toxic exposures can result in death of immune cells, but even small exposures can interrupt immune cell signaling and cell recognition as well as changes in cell proliferation (Corsini, Sokooti, Galli, Moretto, & Colosio, 2013). Surprisingly few studies have documented chemically induced immune suppression that results in a greater number of infections (Person & Whalen, 2010; Phillips, 2000). This is likely the result of non overlapping spheres of research and a remarkably wide divide between the infectious disease and environmental exposure academic communities. Nevertheless there is clear evidence of the immunotoxic effects of many chemicals.

The hypothesized combined effect of both prenatal immune damage and exogenous insults (early life microbial and chemical) to immune function results in an inadequate immune response. This causes further immune dysregulation thereby inducing more intense and higher frequency infections. This is supported by two separate large studies in the United Kingdom and Taiwan have shown that children who go on to develop leukemia have had more medically diagnosed infections than children that do not develop leukemia (Chang et al., 2012; Crouch et al., 2012). The nature of these infections is broad with many groups of infections having been associated.

The scope of the human virome is poorly understood (Delwart, 2013), yet we do know that humans chronically harbor many common viruses. These infections may play critical roles in the development of ALL. We recently observed differences in viral communities in children with ALL during a comprehensive series of viral discovery projects identifying viral sequences recovered from both particle isolation plus 454 sequencing and illumine based whole transcriptome viral discovery. We found more identifiable viruses in children with ALL versus with acute myeloid leukemia’s (AML) (where AML’s have no suspected infectious etiology, yet are immune suppressed). Intriguingly, this excess of infection seems to be driven by double stranded DNA viruses of the Herpes family. Additionally, a differential expression of human endogenous retroviruses (HERV) was also observed, where
ALL’s are producing more HERV’s than AML. Intriguingly, some herpesviruses have been shown to up regulate HERV expression making detangling of the herpes and HERV observation difficult. Some of the infections discovered are common, yet under times of immune suppression, can become over active, potentially suggesting reverse causality. Furthermore there are many unknown sequences actively being transcribed within leukemia samples, these so called ‘dark sequences’ require further study to understand. But a single causal transforming virus has yet to be ruled out in the etiology of ALL.

With a normally functioning immune system, the effect of these viruses would be clinically insignificant. But in the case of children with compromised immune response, we hypothesize that the constant viral challenge, and higher infection load facilitates up-regulation of innate immune response in an attempt to fill the systemic immunologic gap left by inadequate humoral immune response. This constant viral infection causes high viral titers that result in acellular loading of viral nucleic acids. Consequently, the presence of intracellular foreign nucleic acids stimulates persistent innate immune response. The signature of this process is apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like (APOBEC) mediated hypermutation (Alexandrov et al., 2013). These are a family of evolutionary conserved proteins that under normal circumstances would provide some protection by mutating both integrating virus and intracellular viral assemblies. In addition to defending against foreign viruses, a principal function of APOBEC’s is to not only fight off foreign viruses but also to keep HERV’s in check (making the observance of up-regulated HERVS in ALL especially interesting). Yet under persistent viral challenge and inadequate humoral regulation APOBEC-induced somatic hypermutation which likely causes collateral human genomic damage and, ultimately, ALL.

3.5 Future Directions:
We have presented a framework that is a combination of molecular and epidemiologic evidence combined with scientific conjecture. This proposed mechanism contains many gaps, yet we see these gaps as opportunities to target research to further understand the etiologic mechanisms of ALL. Many questions remain:
- Is a specific in utero infection or chemical exposure more common in children that go on to develop leukemia?
- Do children who go on to develop leukemia, have less functional B-cells? Produce more B-cells prior to onset of leukemia?
- Is a specific virus with a tropism for B-cells present in all ALL’s?
- Or is there a suite of viruses with a tropism for B-cells present in ALL’s?
- Do these viruses infect the child prior to the initiation of disease?
- Are APOBEC proteins up regulated in children prior to developing ALL?
- Are other self mutation pathways involved in the etiology of ALL?
- Are germline abnormalities resulting is altered immune function responsible for some risk of ALL?
- Does this framework apply to other cancers?
I look forward to answering these questions.
References


10.1093/ije/dys113


lymphoblastic leukemia in identical twins. *Leukemia, 17*(11), 2202-2206. doi: 10.1038/sj.leu.2403101


DOI 10.1371/journal.pone.0050937


DOI 10.1126/scitranslmed.3000322


Doi 10.1371/Journal.Ppat.1000163


