Genome-wide single nucleotide polymorphism genotyping for identification of copy number variation and allelic imbalance affecting drug response in acute childhood leukemia

Jessica Nordlund

Degree project in biology, 2007
Examensarbete i biologi, 20 p, 2007
Biology Education Centre and the Department of Medical Sciences, Uppsala University
Supervisors: Ann-Christine Syvänen and Lili Milani
SUMMARY

Acute lymphoblastic leukemia (ALL) is the most prevalent cancer in children, accounting for over 30% of newly diagnosed childhood cancers every year. With advances in cancer treatments, the event-free five-year survival rate is nearly 80%. Yet, 20% of patients still succumb to resistant or recurrent disease. ALL is a heterogenic disease that involves many chromosomal aberrations that are characteristic of leukemic cells. Despite the great efforts leading to the discovery of a number of genetic leukemic subclasses, the aberrations alone do not appear to be definitive of a patient’s prognosis nor response to drug treatment.

Recent developments in genome-wide single nucleotide polymorphism (SNP) arrays have made it possible to elucidate chromosomal aberrations, copy number alterations in key multi-drug resistant genes, and SNPs associated with patient relapse. Sixty-four matched DNA and cDNA (RNA) samples taken from children with acute lymphoblastic leukemia were genotyped using the Human NS-12 BeadChip from Illumina, which provides information for over 13,000 genotypes per sample. Data analyses revealed patterns in chromosomal aberrations, allelic imbalance in ATP-binding cassette (ABC) transporter and cytochrome P450 (CYP) genes, and an association between a SNP on chromosome 3 and patient relapse (p<0.05).

These data demonstrate many advantages of high resolution SNP genotyping technology used in this approach to uncover a number of different aspects of ALL. These advantages include accurate detection of genomic aberrations, determination of SNP allele frequency in cases and controls, and presence of allelic imbalance from blood and bone marrow samples. Pairing a complex disease such as ALL with a genotyping system capable of uncovering underlying genomic relationships may provide the clues to discovering the phenotypic variation in drug response.
INTRODUCTION

Acute Lymphoblastic Leukemia (ALL) is a complex disease in which lymphoblast progenitor cells fail to develop normally into mature white blood cells (Figure 1). This disease is defined by the uncontrolled expansion of hematopoietic stem cells that do not differentiate into normally functioning blood cells (Sawyers et al. 1991). This uncontrolled proliferation of leukemic cells causes accumulation of dysfunctional white blood cells, which interferes with the production of other types of healthy cells produced in the bone marrow. Children with ALL are more susceptible to infection due to the inability of white blood cells to perform normal functions. The buildup of leukemic cells in the bone marrow and release of these abnormal cells into the blood stream causes patients to bleed easily and become anemic due to decreased platelet and red blood cell production. ALL in children is an acute disease, meaning that the disease progresses rapidly. At the time of diagnosis, many patients have >95% leukemic cells in their blood and bone marrow. Once diagnosed, ALL patients need immediate treatment with chemotherapeutic drugs. Diagnosis at later stages of the disease (higher white blood cell counts) puts the patient in a higher risk group and generally infers a poorer long term outcome (Carroll et al. 2003 and Pui et al. 2004).

Figure 1. Differentiation of hematopoietic stem cells. In the bone marrow, hematopoietic stem cells differentiate into the myeloid and lymphoid progenitor lines. Acute lymphoblastic leukemia occurs when lymphoid cells (either B or T cell) fail to develop normally, causing overproduction and accumulation in the bone marrow, which affects normal production of other bone marrow derived cells.
The exact cause of childhood ALL is unknown, but it is believed to stem from various environmental and genetic factors. Leukemia has been linked to in utero exposure to DNA damaging drugs, herbal medications, pesticides, and viral infections. Although poorly understood, gender may also be a factor (Shuster et al. 1998). Previous studies report that male children have a slightly worse prognosis than females (Pui et al. 1999).

Leukemic cells frequently accumulate various genomic aberrations, some of which affect the copy number (prevalence) of an allele. One effect of these aberrations is known as loss of heterozygosity (LOH). LOH occurs when one of the parental chromosomes has been lost and no longer contributes to the genome, leaving only one functional parental chromosome (Figure 2B). LOH can also be copy neutral. This occurs when there are still two copies of the chromosome, but both copies are from the same parent (uniparental disomy) (Pan et al. 2005). LOH is important in disease genetics. If a heterozygous individual who has one normal copy and one defective copy of a tumor suppressor gene loses the normal copy, via deletion or mitotic recombination events, LOH occurs. This leaves the individual susceptible to disease (Figure 2) (Wouters et al. 2007).

![Figure 2: Loss of heterozygosity](image)

(A) An example of an individual heterozygous for a tumor suppressor gene. (B) There are a number of different mechanisms that result in LOH. Regardless of the mechanism, each individual depicted retains only the defective copy of the gene.
ALL is generally divided into two groups, B cell ALL (B-ALL) and T cell ALL (T-ALL), based on the leukemic progenitor cell type. B-ALL can be further divided into established genetic sub-types based on the genomic aberrations accumulated in leukemic cells. Approximately one third of acute lymphoblastic leukemia patients show hyperdiploidy, i.e. more than 50 chromosomes in the leukemia cells (Carroll et al. 2003). Other common aberrations include hypodiploid (<46 chromosomes) and four well known subgroups where specific sections of chromosomes are translocated, which disrupts normal function.

A number of recurrent translocations commonly seen in ALL can be used to predict outcome and identify patients who need more rigorous treatment. For example, patients with a translocation between chromosome 12 and 21 (t12;21) have a better prognosis (the six year survival rate is almost 90%) than children with translocations involving the mixed lineage leukemia gene (MLL) on chromosome 11q23 (Armstrong and Look 2005). The ALL subtype involving translocations between chromosomes 9 and 22 (t9;22) which result in a BCR-ABL gene fusion, is also associated with a poorer prognosis. The five year survival rate for BCR-ABL gene fusion patients is around 29% (Pui et al. 2004). The fourth translocation subgroup involves a translocation between chromosomes 1 and 19 (t1;19), which results in gene fusions with oncogenic potential (Privitera et al. 1992).

With recent advances in modern medicine the complete remission rate for childhood ALL is reaching 80%, but despite these improvements approximately 20% still do not respond to treatment (Silverman et al. 2001). Anti-leukemia drugs have a narrow functional range, in addition unnecessary side effects and patient death from drug toxicity are of concern (Pui et al. 2002, Carroll et al. 2003). Adverse reactions to therapeutic drugs are ranked between the 4th and 6th leading cause of death in the United States (Nebert 1999). In order to limit adverse drug reactions and provide patients with precise treatment there is a need for more augmented therapy geared towards the individual. This is especially relevant in ALL where there is an enormous diversity in genetic aberrations and phenotypes between patients classified in the same subtype. Identifying genes that cause drug resistance, investigating the underlying mechanisms, and creating therapies that target resistant cancer are the pinnacle avenues of current research.

**The Human Genome**

The human genome contains the complete set of information responsible for generating phenotype, which differs in every individual. The human genome contains over 3.2 billion base pairs, encoding more than 20,000 genes which are organized into chromosomes (Human Genome Sequencing 2004). Under normal conditions, humans have 46 chromosomes, typical somatic cells have two copies of chromosomes 1-22 plus
two sex chromosomes (two X chromosomes for a female and one X and one Y for a male). One chromosome is inherited from the mother and one chromosome is inherited from the father. The maternal and paternal chromosomes that form pairs are referred to as homologs. For any given gene, there are two copies or alleles. If the gene on the maternal chromosome is identical to the same gene on the paternal chromosome, the individual is homozygous for that allele. If the genes are different between the two chromosomes, the individual is considered heterozygous. Although each individual diploid organism can only have two alleles, there can be many different alleles for a gene in a population.

As a result of two major efforts, the Human Genome Project and the Human Haplotype Mapping Project, the complete sequence of the human genome and much of its variation is now known and available in public databases (Lander et al. 2001, Consortium 2005). In the human genome there are millions of single nucleotide polymorphisms (SNPs). A SNP is defined as a change in a single base in the genome occurring with a frequency above 1% in the population (Kruglyak and Nickerson 2001). Approximately one SNP occurs in every 300 base pairs in the human genome. Because of this high frequency, SNPs are the most abundant form of DNA polymorphisms in humans (Mohr et al. 2002).

Although there are many SNPs spread throughout the genome, only a small fraction of these are nonsilent (nonsynononumus) SNPs found in transcribed regions, or regulatory SNPs (rSNPs) found in regulatory regions. A nonsynononumus SNP causes a base change in a codon that alters the corresponding amino acid, which may alter the protein function. These small differences between individuals cause normal phenotypic variation, but some nonconservative changes may have important clinical consequences.

Identifying the genes responsible for complex diseases has been difficult because each gene only makes a small contribution to the whole phenotype of multifactorial disease. The principal of genetic association studies is to genotype numerous cases and controls and then observe different patterns of SNPs in the two groups (Risch 2000). If SNPs are found highly associated with a disease, one can deduce that nearby the SNP must be an important gene for this disease. In this way whole-genome studies offer a powerful approach by finding SNPs that may help elucidate the causes of human multifactorial disease by uncovering the causal genes that when added together explain the disease phenotype. By scanning the entire genome this approach has the power to uncover genes not previously implicated in disease (Hudson 2003).
What is a SNP?

A single nucleotide difference at one base differing between individuals or homozygous chromosomes

How can a SNP be used to predict medical responses?

- Patients who respond/achieve clinical remission
- Patients who relapse/fail to achieve remission
- Predictive of response
- Predictive of non-response

Figure 3; Single Nucleotide Polymorphisms. Scheme adapted from Roses, 2000.

From Genotype to Phenotype

With the technical improvements in assessing genetic variation, specifically SNPs, the association between genotype and disease susceptibility, outcome, and response to treatment is being analyzed. For example, recently such an approach has identified SNPs in promoter regions of the G1/S checkpoint regulators of a cell that increase susceptibility to ALL (Healy et al. 2007) and the BRCA1/2 genes involved in breast cancer progression (Puputti et al. 2006). It is believed that identifying SNPs in genes that influence the effect of chemotherapeutic drugs may facilitate the understanding of the variety in drug response phenotypes and the multiple genes that are involved (Carroll et al. 2003) (Figure 3). Protein levels regulate numerous pathways in the cell, alteration of regulatory elements of biological processes may influence disease or drug response phenotype. Determining which variations may be important for predicting responses to treatment may be the key to individualizing therapy in order to achieve the best outcome for the patients (Roses 2000).

SNP Genotyping in Pharmacogenetics
SNP microarrays can be used to detect patterns in a wide variety of genomic events such as duplications, monoallelic and biallelic deletions, and allelic imbalance. Novel microarray technologies may provide a key to identify SNPs associated with complex diseases and offer new insight for pharmacogenetic studies (Syvanen 2001). Pharmacogenetics is the study of how genetic composition influences a patient’s response to drugs (Roses 2000). When a drug is introduced into the body, its effectivity is affected by uptake into the cell, binding, metabolism and excretion. The majority of differences in responses appear to be due to alterations in protein binding, transporters and affinity (Nebert 1999).

Complex SNP microarrays may be used to correlate patients that are non-responsive vs responsive to treatments and select different sets of SNPs associated with the patient’s phenotype. Using SNPs as a guide, the genes responsible for phenotype may be discovered from whole-genome scans, which may provide information about drug metabolism genes previously unknown to be associated with a particular function. Discovery of such informative SNPs could serve to produce specific SNP arrays to create pharmacogenetic profiles, providing more personal treatment options. In the future, pharmacogenetic profiling may be used to determine whether an individual is likely to respond to a specific medicine, gearing towards more individualized therapy to hopefully reduce the number of drug related side affects (Roses 2000).
Allelic Imbalance

With the recent advances in microarray technology and expanding genome-wide studies, there is substantial new evidence to suggest that allelic imbalance (AI) of two loci of a gene in an individual may be a frequent event (Lo et al. 2003, Pant et al. 2006). AI can be detected in heterozygous SNPs, where the RNA (cDNA) level for each SNP allele is compared to the corresponding DNA from the same sample. When the allele ratio in RNA (cDNA) differs from the allele ratio of 1:1 in DNA it is considered as AI (Pastinen and Hudson 2004) (Figure 4).
A major advantage of using the AI approach is that the levels of two alleles are simultaneously measured relative to each other, in the same sample. Unlike studies using total RNA expression levels of all samples, this approach provides an internal control for the environmental factors that may cause differences between samples (Milani et al. 2007). With the use of whole-genome studies, detecting AI in genes may be used as a guide to identify SNPs with an association to drug response by identifying genes associated with disease or drug response. Detection of AI can be used as a tool to identify genes with regulatory SNPs that in turn are the cause of imbalanced expression levels (Hudson 2003).

SNPs in non-coding regions such as promoters, enhancers, and other regulatory regions may either increase or decrease the expression of a particular gene by altering transcription factor binding sites, other gene regulation sites upstream, or disrupting microRNA binding sites downstream the gene (Wasserman et al. 2000, Pastinen and Hudson 2004). A recent study demonstrated the use of detecting SNPs with AI in coding regions and using AI as an indicator of noncoding regulatory SNPs in a panel of cancer cell lines. The SNPs identified in gene regulatory regions were determined to affect the protein binding interaction of transcription factors, thus affecting the expression of the gene (Milani et al. 2007). Current opinion is that alterations in non-coding regions are affecting multi-factorial disease phenotype. These functional polymorphisms may explain some of the variation in complex diseases, rather than polymorphisms in the coding region of the gene (Pastinen and Hudson 2004).

**Genome-wide SNP Genotyping**

The genotyping of ALL patient samples was performed using a novel genotyping system from Illumina Inc. Illumina’s NS-12 BeadArrays allows for genotyping approximately 14,000 SNPs in approximately 8,000 genes per sample using technology based on an established genome-wide panel of SNPs that are genotyped using allele-specific primer extension (Gunderson et al. 2005). The SNPs in the NS-12 panel were selected by Illumina, via the screening of public databases for all well defined non-synonomous SNPs in the human genome. The final NS-12 panel consists of 11,600 non-synonymous SNPs and 870 tag-SNPs in the Major Histocompatibility Complex (MHC) region. The tag-SNPs are highly informative SNPs with high linkage disequilibrium, which makes it possible to identify genetic variation without genotyping very many SNPs. SNPs in linkage disequilibrium are classified by having non-random association between two or more loci (Devlin and Risch 1995).
DNA and cDNA are amplified by multiple displacement amplification (MDA), a process that employs multiple hexamer primers to amplify the template to have a suitable amount for genotyping. Many random hexamer primers anneal to the template strand and the highly processive Φ DNA polymerase extends the primers, while new primers bind and extend from newly amplified strands (Figure 5). MDA requires only a couple nanograms of input DNA, and amplifies greater than 10 kilobase products. MDA overcomes some of the error caused by traditional polymerase chain reaction (PCR). PCR has been a limitation in whole genome amplification because of the complexities that arise from primer-primer interactions and sequence dependent amplification that occurs when multiplexing (Syvänen 2005). The resulting intricate maze of MDA

Figure 5: Schematic representation of multiple displacement amplification (MDA). The original DNA (cDNA) sequence is indicated in black, the newly synthesized strands are in red, and hexamer primers are indicated by dotted arrows. (1) Random hexamer primers anneal to DNA (cDNA) and the processive phage polymerase extends them in the 3’ direction. (2) Phage polymerase has strand displacement abilities, which displaces the 5’ end of the upstream strand with the newly synthesized strand. (3) Hexamer primers anneal to newly synthesized strands. (4) A hyperbranched network is formed as the reaction proceeds, creating numerous copies of the original DNA. Figure scheme adapted from Lovmar and Syvänen, 2006.
product is fragmented and annealed to immobilized allele specific primers on a beadarray.

The allele specific primer extension reaction is precise, and can resolve the relative amount of two SNP alleles present as a mixture in a heterozygous sample by using one probe for each SNP allele. In case of a perfect match the primer is extended, in case of a mismatch the primer is not extended (Gunderson et al. 2005) (Figure 6). The primers are extended with hapten-labeled nucleotides, which the polymerase is able to incorporate into the sequence (Steemers et. al. 2006).

Figure 6: Allele specific primer extension. Sample DNA hybridizes to 50 base pair long probe sequences with the SNP allele at the 3’ end. Allele specific primer extension occurs when the sample DNA matches the SNP allele specific probe. (A) If there is a match (T-A), polymerase can extend the sequence with detectable nucleotides. (B) If there is not a match (G-A) at the 3’ end of the probe the polymerase will not extend the sequence and no signal will be generated. This figure was provided with permission from Illumina, Inc.
The goal for using this type of whole genome study is to find the SNPs responsible for altering gene function, which may cause susceptibility to a disease or drug response phenotype. It is likely that polymorphisms may account for a large percentage of variation in drug response (both efficacy and toxicity). These polymorphisms can be used to focus on specific loci and determine the genetic structure of the area (Nebert, 1999).

**Copy Number Variation and Loss of Heterozygosity**

Illumina’s SNP genotyping system allows for flexibility in data analysis. There are different aspects of the genome that can be analyzed using the Beadstudio software from Illumina. Genome profiling for chromosomal aberrations such as copy number variation (duplication and loss of heterozygosity) is becoming essential in cancer biology as it is believed that variation in the copy number of genes may play an important role in cancer progression and variation in disease phenotypes (Wouters et al. 2007). To identify sections of a chromosome or an entire chromosome that has lost heterozygosity, the visualization tool, Illumina Genome Viewer (IGV) within the Beadstudio 3.0 software (Illumina) was used to visualize these abnormalities (Figure 7A).

*Figure 7; Chromosomal aberrations in cancer cells.* The blue color represents the paternal chromosome, pink represents the maternal chromosome. (A) Loss of heterozygosity by complete loss of one chromosome or the loss of a portion of the chromosome. (B) Copy number variation is normally seen in cancer cells, the entire chromosome or just a portion of the chromosome can be duplicated.
The IGV tool can also be used to analyze for occurrence of duplications, which occur when there are three homologous chromosomes, a common occurrence in ALL cells (Figure 7B). Using this tool, heterozygous alleles are easily detected in each patient, chromosome by chromosome, enabling analysis of copy number variation by identification of whole chromosome duplications, partial chromosome duplications and regions containing deletions resulting in LOH.

Two units were quantitatively assessed with the IGV tool; log R ratio and B allele frequency. The log R ratio gives an indirect measure of copy number of each SNP by plotting the ratio of observed to expected hybridization intensity. B allele frequency plots the proportion of times an allele is called A or B at each genotype: thus the expected ratios are 1.0 (B/B), 0.5 (A/B) and 0.0 (A/A). These two statistics allow visualization of copy number changes, and loss of heterozygosity (Gunderson et al. 2005). Thus, the IGV software plots the signal from the two alleles for each SNP along by position on the chromosome, creating a visual picture of duplications and LOH in the chromosome.

Approximately 400 patient samples from children with ALL were available for the project. These samples have been collected since 1996 in a project by the Nordic Oncology and Hematology Organization coordinated by Prof. Gudmar Lönnerholm at Women’s and Children’s Health at Akademiska Sjukhuset, Uppsala. Phenotypic data for each patient sample has been thoroughly recorded since sample collection (time of diagnosis). Known phenotypes known include survival rate, ALL subtype (B or T cell), response to the Nordic Society for Pediatric Hematology and Oncology’s (NOPHO) standard treatment with 3 anti-cancer drugs, and in vitro drug sensitivity to 10 anti-cancer drugs (Frost et al. 2003). The clinical outcome is well known from yearly follow-up, with recording of data in the common Nordic childhood leukemia registry.

**Aim**

The aim of this research project was to use SNP microarrays to analyze copy number variation and identify allelic imbalance of SNPs in a large set of genes in ALL patient samples. Using the data generated from genotyping ALL patient samples on genome-wide SNP arrays, I focused on determining if this approach can be used to detect a range of genomic aberrations, uncover SNPs associated with cases compared to controls, and determine if allelic imbalance occurs in two gene families involved in drug metabolism. This type of assay has demonstrated to be a useful tool to begin uncovering mechanisms of drug response in ALL.
RESULTS

This work is part of a larger study using imbalanced allelic expression as a tool to identify SNPs in regulatory regions of genes affecting gene expression and investigate the relationship between regulatory SNPs and drug response in children with ALL. This larger effort is based on the genotyping SNPs in a unique collection of over 700 Nordic ALL patient samples. Here, the first of the samples were genotyped.

Prior to genotyping, DNA and RNA samples were generated and the preparatory procedures (RNA and DNA extraction, RNA quality analysis, and cDNA preparation) were functionally established. Of 370 ALL patient samples available when the project began, 217 were selected (>2 million cells) for simultaneous RNA/DNA extraction, DNA alone was extracted from the samples containing less than 2 million cells. Of the 217 RNA/DNA sample pairs, approximately 73% (158) contained intact total RNA, and thus were available for genotyping and subsequent allelic imbalance analysis. A collection of 64 high quality samples was carefully selected for initial genotyping. RNA was reverse transcribed into cDNA and matched DNA and cDNA samples were genotyped. In total, 13,917 SNPs were genotyped per sample with a 97% success rate, providing genotype information from 64 children with ALL, thus resulting in data for over 800K genotypes. One third of the samples genotyped were taken from peripheral blood and the remaining samples were taken from bone marrow. No difference in background or variation between the replicates was seen in either cell source (data not shown).

**Cytogenetic Analysis by Genotyping DNA Samples**

The chromosomal aberrations that are commonly seen in leukemia, occur only in cancerous cells, and are not typically seen in non-leukemic somatic cells. Inherent properties of leukemic cells, such as the many duplications and LOH events, can result in a drastic difference in appearance from normal somatic cells. Due to this difference, some variation in the sample quality was observed.
Most samples had clear separation between the heterozygous and homozygous SNP genotypes. These samples with little background were most likely composed of >95% leukemic cells, while the more noisy (higher background) samples are more likely to contain a lower proportion of leukemic cells (Figure 8 A,B). The background noise caused by the presence of non-leukemic, healthy cells demonstrates the difference in genetic content between normal and cancer cells.

Duplication and LOH events were determined by visual inspection of each chromosome in each patient sample. An example of an entire chromosome 6 duplication and complete LOH on chromosome 7 are shown in Figure 8 C, D. Duplication events were identified by the presence of two groups of heterozygous SNPs at different allele frequencies. Entire chromosomal regions as well as partial duplications were observed. LOH was identified by the lack of heterozygous SNP clusters at 0.5 B allele frequency. No heterozygous SNPs indicated that one chromosome is either lost or there are two identical copies of the same chromosome, thus no heterozygous loci were observed.
Table 1: Frequency of genomic aberrations in 64 ALL samples by chromosome

<table>
<thead>
<tr>
<th>Chromosome number</th>
<th>Duplications</th>
<th>LOH</th>
<th>Total aberrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.22</td>
<td>0.00</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>0.06</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>6</td>
<td>0.31</td>
<td>0.00</td>
<td>0.31</td>
</tr>
<tr>
<td>7</td>
<td>0.02</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>8</td>
<td>0.09</td>
<td>0.00</td>
<td>0.09</td>
</tr>
<tr>
<td>9</td>
<td>0.03</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>10</td>
<td>0.30</td>
<td>0.02</td>
<td>0.31</td>
</tr>
<tr>
<td>11</td>
<td>0.05</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>12</td>
<td>0.06</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>13</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>14</td>
<td>0.31</td>
<td>0.00</td>
<td>0.31</td>
</tr>
<tr>
<td>15</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>16</td>
<td>0.03</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>17</td>
<td>0.27</td>
<td>0.02</td>
<td>0.28</td>
</tr>
<tr>
<td>18</td>
<td>0.20</td>
<td>0.00</td>
<td>0.20</td>
</tr>
<tr>
<td>19</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>20</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>21</td>
<td>0.17</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>22</td>
<td>0.06</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td>X</td>
<td>0.16</td>
<td>0.00</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Frequencies indicated in bold are significantly different. There was significant variation in the mean number of duplications on chromosomes 4, 6, 10, 14, 17, 18, 21 and X (analysis of variance p<0.0001). There was significantly more LOH on chromosome 9 (analysis of variance p<0.0001).

The 64 patient samples were analyzed with the IGV tool in BeadStudio 3.0 (Illumina) for each chromosome in triplicate. The overall frequencies of genomic alterations detected by the SNP array IGV analysis are summarized in Table 1. From the entire set of 64 patient samples, a total of 1,472 chromosomes were analyzed. The same pattern of commonly duplicated chromosomes was seen throughout the subgroups of the entire set of 64 patients. The most commonly duplicated chromosomes were 4, 6, 10, 14, 17, 18, 21, and X (p<0.0001) and chromosome 10 was the only chromosome with a significantly greater occurrence of LOH (p<0.0001).
### Table 2: Frequency of genomic duplications and loss of heterozygosity in ALL patients

<table>
<thead>
<tr>
<th>Group</th>
<th>Subtype</th>
<th>N</th>
<th>Duplication (mean ± s.d.)</th>
<th>LOH (mean ± s.d.)</th>
<th>Total (mean ± s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-ALL</td>
<td>Hyperdiploidy</td>
<td>20</td>
<td>5.65 ± 2.60</td>
<td>0.45 ± 0.60</td>
<td>6.10 ± 2.75</td>
</tr>
<tr>
<td>B-ALL</td>
<td>t(1;19)</td>
<td>2</td>
<td>0.00 ± -</td>
<td>0.00 ± -</td>
<td>0.00 ± -</td>
</tr>
<tr>
<td>B-ALL</td>
<td>t(12;21)</td>
<td>8</td>
<td>1.13 ± 1.46</td>
<td>0.38 ± 0.52</td>
<td>1.50 ± 1.31</td>
</tr>
<tr>
<td>B-ALL</td>
<td>MLL</td>
<td>4</td>
<td>0.50 ± 0.58</td>
<td>0.50 ± 0.58</td>
<td>1.00 ± 0.82</td>
</tr>
<tr>
<td>B-ALL</td>
<td>BCR-ABL1 t(9;22)</td>
<td>4</td>
<td>4.75 ± 4.92</td>
<td>0.75 ± 0.96</td>
<td>5.50 ± 5.80</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Hypodiploidy</td>
<td>1</td>
<td>0.00 ± -</td>
<td>1.00 ± -</td>
<td>1.00 ± -</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Other</td>
<td>13</td>
<td>1.00 ± 0.82</td>
<td>0.77 ± 1.09</td>
<td>1.77 ± 1.54</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Normal</td>
<td>5</td>
<td>0.40 ± 0.55</td>
<td>0.40 ± 0.55</td>
<td>0.80 ± 0.84</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Total B-ALL</td>
<td>57</td>
<td>2.77 ± 3.11</td>
<td>0.53 ± 0.73</td>
<td>3.30 ± 3.28</td>
</tr>
<tr>
<td>T-ALL</td>
<td></td>
<td>7</td>
<td>0.29 ± 0.49</td>
<td>0.86 ± 0.69</td>
<td>1.14 ± 0.90</td>
</tr>
<tr>
<td>All cases</td>
<td></td>
<td>64</td>
<td>2.50 ± 3.04</td>
<td>0.56 ± 0.73</td>
<td>3.06 ± 3.18</td>
</tr>
</tbody>
</table>

The total number of aberrations and type of aberration (duplication or LOH) were determined by visual inspection using IGV software and averages and standard deviation (s.d.) from the mean calculated by subgroup. There is a significant difference in the mean number of duplications (analysis of variance p<0.0001) between the B-ALL subtypes. B-ALL: B cell acute lymphoblastic leukemia. T-ALL: T cell acute lymphoblastic leukemia.

Duplication and LOH were determined by visual inspection of aberrations spanning entire chromosomes, chromosomal arms and large segments. Each duplication and LOH event was recorded and basic statistics were calculated to determine if there were any differences between the ALL subtypes (Table 2). There was significant difference in occurrence of duplications in the sub groups (p<0.0001), but no significant different was seen between the different subgroups in regards to LOH. The majority of B-ALL cases were hyperdiploid, which was the only B-ALL subgroup with a significantly higher number of aberrations (p<0.05) upon analysis of variance testing, which was expected. On average these patients had a total of five duplication events, the most commonly duplicated chromosomes were 4, 6, 10, 14, 17, 18, 21, and X. Most of the B-ALL subgroups had noticeably high standard deviations in regards to copy number and some of these also demonstrated hypodiploidy in addition to the translocation by which they were classified. The high standard deviations could be due to the low sample size or a reflection of the inherent properties of ALL.
LOH analysis for each chromosome was performed using IGV to visualize the DNA genotypes. One distinct pattern was identified. In the entire sample set (regardless of subtype), chromosome 9 had a significantly greater degree of LOH than any other chromosome. Further examination into the high frequency of LOH on chromosome 9 revealed that the most frequently deleted regions were on the short arm (9p). Chromosome 9p contains genes that encode proteins that are important to B cell differentiation and have been associated with cancer progression. 9p21.3 is a region that encodes \textit{CDKN2A}, cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4). \textit{CDKN2A} is a known important tumor suppressor gene that encodes at least three splice variants. Two variants function as CDK4 kinase inhibitors, while the third is known as a stabilizer of p53. These variants share a common function in cell cycle G1 control (Zhu \textit{et al.} 2006). In addition, this region contains Paired Box 5a (\textit{PAX5}), a gene located at 9p13, which encodes a member of the paired box family of transcription factors which is important in B cell lineage activation at the early stage of B cell differentiation (Nutt \textit{et al.} 1999).

\textbf{SNP Allele Frequency in Cases vs. Controls}

The whole genome association analysis tool, PLINK (Purcell \textit{et al.}, in press) was used to analyze 10,405 SNPs in a case-control comparison. PLINK found one candidate SNP with a minor allele frequency differing between the relapse patients (case) and patients in complete remission (control) (p<0.05 after Bonferroni correction for multiple testing). The minor allele frequency of the candidate SNP was 36% in the relapse patients compared to 4% in continuous clinical remission (CCR) patients. A SNP panel designed to determine variation in the human genome (from Perlegen) reports the minor allele frequency for this SNP to be 8% in the European population. This C/T SNP is located on chromosome 3, where C is the ancestral allele and T is the minor allele. The minor allele (T) is a nonsynonymous SNP that changes an amino acid in the encoded protein from arginine to cysteine. This change in amino acid switches from a polar, basic amino acid, which binds negative charges to a neutral amino acid. This SNP was found on a candidate gene that may have an unknown function in drug response phenotype.
Figure 9: Example of an allelic imbalance ratio plot. The allelic ratio in DNA is plotted on the X axis and the allelic ratio in RNA is plotted on the Y axis. Ratios are calculated by the signal from the A allele divided by the signal from the A and B alleles combined (A/(A+B)). Homozygous B allele SNPs are in the lower left hand corner (0,0), heterozygous SNPs are located in the center (0.5, 0.5) and homzygous A allele SNPs are located in the upper right hand corner (1,1). SNPs indicated in red and green show significant imbalance. Over expressed A alleles are located above the heterozygous cluster and over expressed B alleles are located below the heterozygous cluster.

Allelic Imbalance in Gene Expression by Genotyping RNA

For a SNP to be informative in detecting AI, at least one individual must be heterozygous for the particular SNP. Using an R script for allelic imbalance, 11,733 informative SNPs were detected in 6,329 genes. Imbalanced expression was observed in 1,963 (17%) informative SNPs, in 1,519 genes in two or more heterozygous individuals. The allelic ratio was calculated by dividing the allelic fraction (A/(A+B)) in RNA with that in corresponding DNA. Imbalanced expression occurs when the RNA/DNA quotient does not equal one (Figure 9).

The AI approach was used to focus on two specific gene families, encoding ABC transporters and CYPs (cytochrome P450). These two families are known to be involved in multi-drug resistance in cancer (Sarkadi et al. 2006). The Illumina NS-12 beadarray contained 87 SNPs in 37 ABC transporter genes and 59 SNPs in 29 CYP genes. Imbalanced allelic expression was observed in 11 of the 37 (30%) ABC genes and in 7 of the 29 (24%) CYP genes.
DISCUSSION

The enormous complexity of childhood ALL is one of the principle barriers to discovering the pathway of leukemogenesis. The diversity in the genetic aberrations and differences in drug response from one patient to another provide little clue to the origin and bases of this disease. The wide variety of abnormalities creates various subgroups with tendencies to experience more positive or negative outcomes, but this is not absolute. Hyperdiploidy has long been known to be associated with a positive response to cancer medication (Williams et al. 1982), but the molecular basis for this observation has not been thoroughly explained. With recent genome wide studies and technological advances clues to the origin of this disease are being uncovered. Identification and classification of the various genetic aberrations and accumulation of large ALL cohorts have brought important insight into leukemia progression. High resolution genetic studies may have the capability of uncovering the reasons why some sub-types of ALLs respond to treatment and others do not (Downing and Mullighan 2006).

Chromosome 9

Abnormalities on chromosome 9p have been associated with negative outcomes for children with ALL (Heerema et al. 1999). A previous whole genome association study found that PAX5 on chromosome 9p is mutated in approximately 31% of ALL cases and is thought to have a dose related effect on pathogenesis of childhood ALL (Mullighan et al. 2007). The data from this current study showed that in 30% (19/64) of ALL cases there was some genomic abnormality, such as duplication or LOH, in this region. Thus essentially replicating the result from Mullighan et al.. Although the Illumina NS-12 SNP array does not have SNP coverage on PAX5, analysis of total aberrations on 9p provide evidence to suggest that chromosome 9p13 and other flanking segments on this chromosome are in fact altered (duplicated or LOH).

Drug Metabolism Genes

A number of ABC transporters are known to contribute to multi-drug resistance, among them ABCB1, ABCC1 and ABCG2 (Sarkadi et al. 2006). Using AI of SNPs as a guide, 11 genes with imbalanced expression including ABCB1, ABCC1, and ABCG2 were identified in the ALL patient samples. ABC over expression is well known to infer drug resistance in leukemia patients. Interestingly, ABCG2 is highly expressed in normal hematopoietic stem cells, but very little in mature blood cells and may be one of the reasons why resistant cancerous stem cells remain after chemotherapy treatment, causing
relapse months to years after treatment (Dean et al. 2005). Imatinib (Glivec), a widely used anti-cancer drug, has recently been shown to be an ABCG2 substrate (Houghton et al. 2004).

Therapeutic drugs are often metabolized by cytochromes 450P (CYP), which belongs to a superfamily of hemoproteins (Miners and Birkett 1998). Using the AI approach seven out of 29 CYP genes covered on the Illumina array demonstrate imbalanced expression. CYP genes encode cytochrome enzymes that are involved in metabolism of various drugs. A recent study in a Swedish cohort found a linkage between CYP2C8 and CYP2C9 allelic variants and relationship to impaired drug metabolism (Yaser et al. 2002). A wide inter-individual variation exists in regards to CYP activity, but the source of this variation is not yet understood. AI in genes such as the CYP3A sub family were previously recorded, and alterations in methylation patterns inconsistent with parental imprinting were observed (Hirota et al. 2004). The AI method has a strong foundation and provides a novel approach to treatment-response evaluation. Using AI as a tool to identify SNPs in genes involved in drug metabolism and export from the cells, may shed light on this complex disease and help uncover disease genotypes which will enable physicians to implement more individually orientated medications.

**Future Prospective**

The success of genotyping 64 ALL patient samples in DNA and RNA (cDNA) showed that this approach has potential to elucidate novel SNPs which may serve as a guide to understanding this complex disease. All of the results must be confirmed in a larger sample set in order to provide more genotypes and samples in the different subclasses, providing more significant results. Once the remaining samples are genotyped, genes displaying AI will be selected and the promoter regions will be sequenced to elucidate regulatory SNPs and methylation patterns that may be causing the imbalance. Finally, all of the genotypes, regulatory SNPs, and methylation patterns will be correlated to genomic aberrations, clinical remission status and drug responses. Considering the various complex levels that can be studied paired with the innate complexity of ALL, implementing a genotyping system equally complex and capable of uncovering genetic polymorphisms should provide the clues to elucidate phenotypic variation in drug response. If this approach is successful, using AI of coding SNPs as a guide to uncover functional polymorphisms can be explained and use to elucidate mechanisms of other complex diseases.
MATERIALS AND METHODS

ALL Patient samples

The collection of patient samples consisted of cells from peripheral blood or bone marrow from approximately 400 patients of northern European descent. All patients could be classified by subtype as having hyperdiploidy, translocations t(1;19), t(12;21), BCR-ABL1 t(9;22), MLL, hypodiploid, other genomic abnormalities, being cytogenetically normal, and T-ALL patients (Table 3). All samples have been stringently recorded for remission status. Of the 64 patients genotyped so far, 51 patients are in complete remission (CCR), 11 have relapsed, and 2 have died from causes other than disease relapse. Other phenotypes recorded include response to NOPHO standard treatment with three anti-cancer drugs, in vitro drug sensitivity to ten anti-cancer drugs and continuous clinical remission status is recorded yearly in the common Nordic Childhood Leukemia Registry.

Table 3; Description of ALL subtypes

<table>
<thead>
<tr>
<th>Group</th>
<th>Subtype</th>
<th>N</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-ALL</td>
<td>HH</td>
<td>20</td>
<td>Hyperdiploid with &gt;50 chromosomes. Pattern of 4,6,10,14,18 and 21. 4 and 10 associated with good prognosis</td>
</tr>
<tr>
<td>B-ALL</td>
<td>t(1;19)</td>
<td>2</td>
<td>Translocation between chromosome 1 and 19</td>
</tr>
<tr>
<td>B-ALL</td>
<td>t(12;21)</td>
<td>8</td>
<td>Translocation between chromosome 12 and 2. Associated with a good prognosis</td>
</tr>
<tr>
<td>B-ALL</td>
<td>MLL</td>
<td>4</td>
<td>Myeloid/lymphoid or mixed lineage leukemia gene (MLL) involve a rearrangement at 11q23</td>
</tr>
<tr>
<td>B-ALL</td>
<td>BCR-ABL1 t(9;22)</td>
<td>4</td>
<td>Translocation between breakpoint cluster region gene (BCR) on chromosome 9 and v-abl Abelson murine leukemia viral oncogene homolog 1(ABL1) on chromosome 22. Results in a BCR-ABL1 hybrid oncogene</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Hypodiploidy</td>
<td>1</td>
<td>Less than 46 chromosomes, generally unfavorable prognosis</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Other</td>
<td>13</td>
<td>Other subclasses of ALL including containing chromosomal abnormalities not covered by the main subgroups</td>
</tr>
<tr>
<td>B-ALL</td>
<td>Normal</td>
<td>5</td>
<td>Chromosomes appear normal (&lt; 1 chromosomal aberration)</td>
</tr>
<tr>
<td>T-ALL</td>
<td>T cell</td>
<td>7</td>
<td>ALL originating from T-cell progenitors</td>
</tr>
</tbody>
</table>
Cell pellets containing from 0.2-8 million cells were stored at -70°C prior to the project. Patient samples were divided into two groups based on the total number of cells in each sample. DNA and RNA were purified from samples containing from 2 to 8 million cells, while DNA alone was purified from samples containing less than 2 million cells.

**DNA and RNA Purification**

DNA and RNA were extracted from 217 patient samples containing over 2 million cells using the AllPrep DNA/RNA Mini Kit (QIAGEN). All procedures were followed according to the manufacturer’s protocol for animal cells. The optional On-Column DNase Digestion protocol was also included to prevent DNA carry over into the RNA samples. DNA concentration was measured with a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies) using 1.5 µl of DNA from each sample for analysis. All purified DNA and RNA were then stored at -70°C.

**DNA Purification**

DNA was purified from 153 samples containing less than 2 million cells with the QIAamp® DNA MiniKit according to the QIAamp DNA Blood MiniKit Handbook (QIAGEN). Cell pellets were first re-suspended in 1% Phosphate-buffered saline (8 g NaCl, 0.2 g KCl, 1.44 g Na2HPO4, and 0.24 g KH2PO4 in 800 ml of RNase-free H2O) to a final volume of 200 µl before adding 20 µl of Qiagen proteinase K and then continuing according to the manufacturer’s Blood and Body Fluid Spin Protocol. DNA was eluted by adding 50 µl of AE buffer to the spin column and incubating for 5 minutes at room temperature before spinning. This was repeated with a second aliquot in order to gain maximum DNA recovery. DNA concentration was measured with a NanoDrop® ND-1000 Spectrophotometer (NanoDrop Technologies) using 1.5 µl of DNA from each sample for analysis. All samples were then stored at -70°C until data analysis.
Figure 10: Electrophoretic analysis of 18s and 28s rRNA. The Y axis shows the level of fluorescent units (FU) and the X axis shows the time in seconds. The peak emerging at 23 seconds in the ladder control, and the peaks appearing at 43 and 50 seconds are 18s and 28s rRNA respectively. (A) RNA sample with high integrity, strong peaks, good 18s to 28s rRNA ratio, and higher fluorescence. (B) An example of a degraded RNA sample, no distinction between 18s and 28s rRNA, high background from 25 to 40 seconds, and lower fluorescence.

RNA Purity and Integrity

The quality of RNA is critically important for cDNA amplification and genotyping assays. Impurities and contamination in the RNA sample can lower yield and reduce the level of transcription according to the Illumina® TotalPrep RNA Amplification Kit protocol (Ambion). To evaluate RNA integrity, 1 µl of each RNA sample was subjected to microfluidics (electrophoretic) analysis using the Agilent® 2100 bioanalyzer and a RNA LabChip®. This assay determines sizes of RNA and identifies samples containing degraded RNA. Degraded samples perform poorly in cDNA synthesis, creating short transcripts that can lack important regions. Figure 10 shows a result of the microfluidics analysis, RNA sample of good quality (A) and a degraded RNA sample unsuitable for cDNA synthesis (B).

To screen for DNA contamination, 1 µl from each RNA sample was amplified by polymerase chain reaction with SmartTaq polymerase (Nexo Lab) and 10µM genomic primers which amplifies a 90bp segment around in the PARP1 gene around exon 9 (PARP1 forward primer: (5'->3') TCTGACATGTTTCTCCACTGG, PARP1-reverse primer: (5'->3') CGAGCATTCCCTGCAGGG, from Integrated DNA Technologies). The PCR products were run on 2.5% agarose gel, 0.5 µg/ml ethidium bromide in 0.5% Tris-Borate-EDTA buffer (54 g of Tris base, 27.5 g of boric acid, 20 ml of 0.5 M EDTA (pH 8.0) in 2 liters H20) for 45 minutes at 150 V and 90 mA. All RNA samples did not contain DNA.
**cDNA Synthesis**

64 RNA samples that passed RNA purity control were reverse transcribed into double stranded cDNA with the Illumina® TotalPrep RNA Amplification Kit (Ambion), providing high quality doublestranded cDNA for array analysis using 1 µg input RNA. The first strand cDNA was synthesized by reverse transcription with T7 Oligo(dT) Primer. The second strand cDNA was synthesized to convert the single stranded cDNA into double stranded cDNA with DNA Polymerase and RNaseH to degrade any remaining single stranded RNA. The last step was to purify the cDNA, to remove excess RNA, enzymes, primers and salts that may interfere with downstream reactions. All steps were carried out following the manufacturer’s protocols. All cDNA was then stored at -70ºC after purification.

**Genome-wide SNP Genotyping**

Paired DNA/cDNA samples were genotyped in triplicate on NS-12 Beadarrays (Illumina). The array can genotype 14,000 SNPs using an allele-specific primer extension assay on beads that are grafted on 3-micron silica beads, which are assembled onto microscope slides. Twelve samples (2 DNA and cDNA sample-pairs in triplicate) were analyzed for the whole SNP panel on one microarray slide. The allele-specific primer extension reaction is specific, and allowed for determination of the relative amount of two SNP alleles that were present as a mixture in a sample by highly precise probes that could discriminate between the two SNP alleles. This was done by using two bead types for the two different SNP alleles, the only difference in the 3’ terminal base, creating allelic discrimination in the polymerase extension step (Gunderson et al. 2005).

All experiments were carried out according to the multi-step manufacturers protocol for the Infinium™ 1 Assay Human NS-12 Multi-Sample BeadChip (Illumina). The first step of the genotyping procedure was whole genome amplification by multiple-displacement amplification (MDA), which used the highly processive Φ29 polymerase to create DNA products which are >10 kb in length (Barker et al. 2004, Gunderson et al. 2005, Lovmar and Syvanen 2006).

The second step was preparation of the MDA product for hybridization by fragmentation, precipitation, and re-suspension. The DNA was then hybridized to capture oligonucleotides on the BeadChip, which are 50 base pairs in length. The third step was allele specific primer extension using two bead types which only differ at the 3’ terminal base at the SNP position. The extension was carried out with labeled nucleotides and signal amplification and then the BeadChip was read on Illumina’s
Figure 11: Three day workflow for genotyping 14,000 SNPs on NS-12 Beadarrays (Illumina). Day 1; multiple displacement amplification (MDA). Day 2; amplified DNA and cDNA are fragmented, precipitated and re-suspended. Samples are added onto the beadchips and hybridize overnight to corresponding allelic probe sequences at 45°C. Day 3; the hybridized samples are extended and stained before imaging on a Beadarray Reader. Figure from Illumina.com with permission.

Beadarray Reader. The genotype of any given SNP was AA, AB, or BB, and was determined by the intensity ratio between the two alleles ($\Theta=2/\pi*\arctan(B/A)$) (Gunderson et al. 2005). See Figure 11 for an overview.
Data Analysis

The SNP genotypes were assigned using the Beadstudio software (Illumina), raw A and raw B allele intensity values were exported and analyzed using an R script written for AI analysis. Illumina’s Genome Viewer was used to visualize chromosomal duplications and deletions (LOH). Associations between relapsed patients and clinical remission patients were performed using gPLINK, a whole genome association analysis toolset (Purcell et al. 2007) (http://pngu.mgh.harvard.edu/purcell/plink/).
ACKNOWLEDGEMENTS

I would first like to thank Ann-Christine Syvänen for welcoming me into the Molecular Medicine group and in trusting me with this exciting project. I would especially like to thank Lili Milani for being a phenomenal supervisor and enormous support throughout. I also would like to thank Anders Lundmark for his support, specifically for writing the R script for allelic imbalance analysis and plotting/visualizing data. Thank you to everyone at the platform as well for help with the Illumina experiments and all the others in the Molecular Medicine group for the advice, inspiration and friendly environment. Finally, a big thank you to my parents, Luam, Suzanne, and Britta for carefully proofreading my thesis.
REFERENCES


Purcell BN, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Mallar J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A toolset for whole genome association and population based linkage analyses. Am J Hum Genet 81:559-575


