Genomic characterization of pediatric acute lymphoblastic leukemia by deep sequencing

CARL MÅRTEN LINDQVIST
Dissertation presented at Uppsala University to be publicly examined in E10:1307-1309, Navet, BMC, Husargatan 3, Uppsala, Uppsala, Friday, 19 February 2016 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Professor Gabor Marth.

Abstract

Acute Lymphoblastic Leukemia (ALL) is the most common cancer in children, with close to 200 cases per year in the Nordic countries. Despite recent advances in modern chemotherapies, 20% of the ALL patients experience a relapse. ALL has traditionally been stratified into subtypes with different risk classification and therapy using large genomic aberrations such as translocations and aneuploidies. In recent years technological advances have enabled the detection of smaller genetic variants, such as point mutations and small insertions/deletions. This thesis focuses on the detection of these smaller variants and their potential impact for ALL.

The present work includes four studies. In the first study we investigated the effects of whole genome amplification and non-indexed pooling strategies to maximize the output of targeted sequencing. We found that whole genome amplified DNA is equivalent to genomic DNA when screening for point mutations in targeted sequencing data. We were able to accurately detect variants in non-indexed pools with up to ten samples. The second study describes further work on non-indexed pools where we pooled samples in an overlapping scheme and identified carriers of rare variants. The third study describes the whole genome and RNA sequencing of four patients with ALL and validated the results in a cohort of 168 additional ALL patients. In the whole genome sequenced patients we found somatic mutations in both known cancer driver-genes (KRAS and NOTCH1) and in putative driver-genes (KMT2D and KIF1B) after analysis of the additional ALL patients. We validated point mutations genome-wide and observed a large number of C>A mutations in one patient, in contrast to C>T mutations that are more common in cancer in general. The fourth study analyzed the same cohort as the third study and expanded the target to 872 genes linked to cancer, ALL or epigenetic regulation recorded in the literature. We found distinctive differences between BCP-ALL and T-ALL both in number and types of mutations. In addition we observed an association between mutations in the Notch signaling pathway and relapse.

These results will have an impact on future studies of cancer, and add another piece to the genetic puzzle of ALL.

Carl Mårten Lindqvist, Department of Medical Sciences, Molecular Medicine, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden.

© Carl Mårten Lindqvist 2016
ISSN 1651-6206
urn:nbn:se:uu:diva-269760 (http://urn.kb.se/resolve?urn=nbn:se:uu:diva-269760)
To my family
This thesis is based on the following papers, which are referred to in the text by their Roman numerals.


* These authors contributed equally to the work
## Contents

### Introduction ................................................................. 11
  Acute Lymphoblastic Leukemia ........................................... 11
  Genetic variation in cancer .............................................. 12
  Heterogeneity in cancer .................................................. 12
  Group testing ............................................................... 13

### Aims .................................................................................. 14

### Materials and Methods ...................................................... 15
  Patient Samples .............................................................. 15
  Technologies ....................................................................... 15
  Variant Calling ................................................................... 16
  Validation ........................................................................... 16
  Pooling ............................................................................... 17
  Identification of driver genes ............................................. 18

### Results ............................................................................. 20
  Paper I: Accurate detection of subclonal single nucleotide variants in
  whole genome amplified and pooled cancer samples using HaloPlex
  target enrichment .......................................................... 20
    Validation ........................................................................ 20
    Whole genome amplified DNA compared to genomic DNA ...... 22
    Non-indexed pooling of DNA ........................................... 22
  Paper II: Identification of somatic variants by targeted sequencing of
  pooled leukemia samples .................................................. 23
    Verification of pooling ................................................... 23
    Sensitivity and Precision ................................................ 24
    Novel variants ............................................................... 24
  Paper III: The mutational landscape in pediatric acute lymphoblastic
  leukemia deciphered by whole genome sequencing ................. 24
    Somatic mutations ......................................................... 25
    Genome-wide mutational patterns ..................................... 25
    Recurrent mutations ...................................................... 26
  Paper IV: Targeted sequencing of cancer genes in pediatric acute
  lymphoblastic leukemia ....................................................... 26
    Somatic mutations in diagnosis and relapse ....................... 26
    Driver genes ..................................................................... 27
    Survival analysis ........................................................... 27
Summarized discussion ................................................................. 28
Acknowledgement ........................................................................ 30
References ..................................................................................... 32
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF</td>
<td>Allele fraction</td>
</tr>
<tr>
<td>ALL</td>
<td>Acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>BCP-ALL</td>
<td>B-cell precursor acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>CNA</td>
<td>Copy number alteration</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>gDNA</td>
<td>Genomic DNA</td>
</tr>
<tr>
<td>Indel</td>
<td>Insertion-deletion</td>
</tr>
<tr>
<td>LOH</td>
<td>Loss of heterozygosity</td>
</tr>
<tr>
<td>NGS</td>
<td>Next generation sequencing</td>
</tr>
<tr>
<td>nsSNV</td>
<td>Non-synonymous single nucleotide variant</td>
</tr>
<tr>
<td>NOPHO</td>
<td>Nordic Society of Pediatric Hematology and Oncology</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>SNV</td>
<td>Single nucleotide variant</td>
</tr>
<tr>
<td>T-ALL</td>
<td>T-cell precursor acute lymphoblastic leukemia</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>WGA</td>
<td>Whole genome amplification</td>
</tr>
<tr>
<td>wgaDNA</td>
<td>Whole genome amplified DNA</td>
</tr>
<tr>
<td>WGS</td>
<td>Whole genome sequencing</td>
</tr>
</tbody>
</table>
Introduction

In 2001 the international Human Genome Project presented a first draft of the sequence of the human genome (1). The sequence consisting of just over 3 billion base pairs is approximately 99.9% similar between individuals. The genome of cancer cells has often numerous differences compared to the genome of healthy cells from the same person, including variation at single bases and larger insertions, deletions or rearrangements. Additionally, a cancer tumor is often genetically diverse, with different cells in the same tumor having different versions of the genomic sequence. New sequencing technologies enable us to examine how the cancer and normal genome differ in a much larger scale than what has been seen before. This thesis covers the use of these sequencing methods to characterize the ALL genome.

Acute Lymphoblastic Leukemia

Acute lymphoblastic leukemia (ALL) is the most common cancer in children, with approximately 200 cases per year in the Nordic countries (2). Despite recent advances in modern chemotherapies, 20% of ALL patients do experience a relapse. ALL patients are generally divided into T-cell and B-cell precursor phenotypes (T-ALL and BCP-ALL), depending on the type of lymphocytes the leukemia cells originate from (3). Patients are further classified by different recurrent somatic aneuploidies and structural variants. These subtypes, together with non-genomic clinical factors such as age and white blood cell count, have been clearly linked to risk of treatment complications or later relapse and are essential factors for the choice of therapy of ALL patients (4). Although there are indications of elements of inheritance in ALL such as different incidence in different populations (5), higher frequency of some SNPs in ALL-cases (6,7) and higher risk if a twin is affected (8), inheritance seems to play a minor role in leukemia (9). Infections have been reported to be involved in development of ALL (10) and might explain assumed inheritance.
Genetic variation in cancer

Genetic predisposition is of various importance in different types of cancer. An early study of retinoblastoma showed a clear pattern of inheritance for certain individuals and became the basis for the multiple hit hypothesis (11). This hypothesis proposes that cancer develops after two or more hits (mutations) (12,13). The first hit might be an inherited variant or somatic mutation. It is worth noting that this hypothesis was mainly based on the distribution of age when the retinoblastoma occurred and in other types of cancer the number of hits might differ. The hypothesis got molecular support with the discovery of germline variants in RB1 associated with familiar cases of retinoblastoma (14). Later studies have revealed inherited variants in e.g. BRCA1/2 as a risk factor for breast cancer (15), and today more than 100 genes have been linked with predisposition to cancer (16). Still the most important variation seems to be somatically acquired mutations (17). Soma-
tically acquired mutations are spread all over the genome but cluster in cer-
tain genes that seem to “drive” the cancer. Theses so-called driver genes can
further be separated into oncogenes (where mutations change the protein to
gain a function that drives the cancer) and tumor suppressor genes (where
mutations inactivate a gene that protects from cancer). Oncogenes tend to
have the same position recurrently mutated in different patients, whereas
tumor suppressor have mutations spread out over the gene in positions that
inactivate the gene (17). Mutations that seem to stem from a random event
and have no functional impact on the progression of the cancer are called
“passenger mutations”.

Mutations in cancer can be of any size from extra chromosomes (aneu-
ploidy) down to the change of a single base, single nucleotide variant (SNV).
In between come small insertions or deletions (indels); larger insertions or
deletions, so-called copy number alterations (CNA); rearrangements, both
within and between chromosomes. Most analyses in this thesis are focused
on the small variants, especially SNVs.

Heterogeneity in cancer

Nowell proposed a model of somatic development of the cancer in a evolu-
tionary pattern originating from one cell (18). This means that different pop-
ulations of cells might co-exist, where some mutations are shared by all cells
and were present in the original cancer cell, and some mutations are only
found in a subset of the cells and have hence occurred at a later stage. An
example of this in a cancer sample can be seen in Figure 1. The model has
later received experimental support in different types of cancer (19) includ-
ing ALL (20). Subclones can be of clinical importance if the mutations they
harbor make them resistant to therapy (21).
Figure 1. Example of heterogeneity in a cancer sample. A mixture of 10 cancer cells and 1 normal cell represent normal contamination of about 9%. All cancer cells have mutation ‘x’, and the cells with no additional mutations belong to the founder clone, which will be near-identical to the original cancer cell. Mutation ‘o’ exists only in 4 cells, hence belongs to a subclone that consists of 40% of the cancer cells.

Group testing

The theory of mixing samples into pools as a way to increase throughput of clinical testing for rare diseases was first proposed by Dorfman in 1943 (22), where testing for syphilis in blood was used as an example. Since then this strategy has been applied in different settings, including genetic studies (23,24). Although the pooling can be performed in many ways, mixing equal amounts of DNA from each patient is the most straightforward design.

Figure 2. Example of overlapped pooling. Patient 1 is the only one present in both pools. Mutations found in these two pools but not in any other pool where patient 2-19 is present can then be attributed to patient 1. Bars to the right represent sequencing reads and red reads originate from patient 1.
Aims

The aim of this thesis was to study the pattern of somatic mutations in ALL patients using next generation sequencing (NGS). Specific topics are as follows:

1. Find mutations relevant to the disease progression of ALL
2. Investigate novel methods to detect mutations in deep sequence data
Materials and Methods

Patient Samples
Diagnostic samples from ALL patients treated at four collaborating centers within the Nordic Society for Pediatric Hematology and Oncology (NOPHO) were available for the studies. Diagnostic bone marrow and/or blood samples from 342 patients treated at any of four centers were analyzed (Uppsala, Umeå, Stockholm and Gothenburg). Matched normal DNA samples were available for 163 of the patients. RNA was available for 4 patients (Uppsala and Stockholm) and an in-house RNA-seq dataset containing 27 BCP-ALL samples and 18 T-ALL samples (Nordlund et al., unpublished data) was used to determine gene expression levels in ALL cells. Both DNA and RNA were extracted from blood cells collected at diagnosis. Four of the patients were selected for whole genome sequencing (WGS) of both leukemic and matched normal sample Table 1. RNA sequencing was performed for the leukemic sample.

Table 1. Patients subjected for Whole genome sequencing.

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Immunophenotype</th>
<th>Genetic subtype</th>
<th>Patient ID in study 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL_458</td>
<td>BCP-ALL</td>
<td>t(12;21)</td>
<td>ALL1</td>
</tr>
<tr>
<td>ALL_559</td>
<td>T-ALL</td>
<td>T-ALL</td>
<td>ALL2</td>
</tr>
<tr>
<td>ALL_707</td>
<td>BCP-ALL</td>
<td>Other</td>
<td>N/A</td>
</tr>
<tr>
<td>ALL_501</td>
<td>BCP-ALL</td>
<td>Normal</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Technologies
The majority of the data in this thesis has been generated using next generation sequencing (NGS) by single-base primer extension (Illumina, USA). By sequencing many short DNA fragments in parallel it is possible to analyze a whole human genome within weeks. The same technology has been used to determine gene expression through RNA sequencing. For sequencing regions of interest with NGS, Haloplex targeted sequencing has been used (Agilent, USA). Genotyping was performed with Illumina microarrays for quality control of the WGS and as a secondary input for calling copy number alteration (CNA). As the genetic material is biobanked and used for other studies not presented in this thesis there might be small amounts available,
especially for samples taken a long time ago. Samples where only small amounts of DNA were available were subjected to whole genome amplification (WGA).

### Variant Calling

Variant calling for single nucleotide variants (SNV) in WGS data is generally performed using a Bayesian interference model taking posterior probabilities in the genome into account when assigning genotypes (25–27). Since the human genome is diploid many variant callers assume pseudo-discrete distribution of allele fractions (AF) from variants: 0, 0.5 and 1 representing no SNP, heterozygote SNP and homozygote SNP. For somatic mutations in leukemia cells this is not a true assumption due to the heterogeneity in a leukemia-sample, consisting of both leukemia and healthy cells. In addition to the heterogeneity many mutations are subclonal and hence not found in all leukemia cells. Further, it is necessary to filter out germline SNPs. We have used two programs adapted for detection of SNVs in cancer genomes with matched normal sample; SomaticSniper (28) and MuTect (29). For variant-calling in the validation cohort we used FreeBayes (30) together with an internal pipeline to distinguish somatic from germline variants.

### Validation

NGS offers massive amounts of data at low cost making it possible to sequence large genomes, but more data leads to more errors. The error rate for Illumina is about one error per thousand bases sequenced (31) and with the large amounts of bases sequenced in WGS of a human genome (~10^{11} bases for 30x coverage) it becomes very important to filter erroneous bases. As the errors are not randomly distributed in the genome it might be hard to filter all of them (32), especially since the vast number of bases in the human genome makes it likely that erroneous bases are included at some positions. Therefore it is important to perform some form of validation.

The majority of variants detected in the present work were validated using high-depth targeted sequencing data. We checked each position in the targeted sequencing data where we had called a somatic SNV in the WGS data, and defined a validated SNV based on hard cut-offs of AF in tumor and normal sample using an internally developed program, as described in study I (33). As this method uses the same chemistry as the original experiment, we have also amplified the region around some variants with polymerase chain reaction (PCR) and sequenced with Sanger sequencing (34). DNA from whole genome amplification (WGA) was used for PCR and Sanger sequencing to minimize the use of the available material. In addition to ex-
Experimental validation, variant calls were also inspected in Integrative Genomics Viewer (35).

Pooling

In study I we evaluated the effect of DNA-pooling on observed allele fractions (AF) for somatic SNVs. We made three different designs of pooling, with DNA from 2, 5 and 10 patients mixed together. Each design contained a WGS patient with known variants, both germline and somatic. All three designs were performed in replicate. In Study II we pooled patients in an overlapped scheme where each patient was included in two pools and the intersection between these two pools consisted only of that patient. An example of this type of pooling is shown in Figure 3 where 6 patients are pooled in 4 pools. Traditionally pooling has been done using a row-column approach where all patients are placed in a matrix and a pool is assigned to each row and column (Figure 4). When a variant is found in one column-pool and one row-pool it can be attributed to the patient in the intersection of these pools (36). The row-column pooling strategy is less flexible and will always require equal or more pools than the overlapped scheme selected by us for the same number of patients.

Figure 3. Example of a complete graph representing overlapped pooling of 6 patients in 4 pools. Each pool (black open circles) contains three patients (filled numbered circles). Lines between two pools are color-coded according to the patient shared between them. For each pair of pools where a rare variant is found, a carrier can be uniquely identified.
Figure 4. Row-column pooling. All patients (filled numbered circles) are placed in a matrix. Pools (black open circles) are assigned to every row and column. Every patient is intersecting and included in one column- and one row-pool. For each pair of column- and row-pool where a rare variant is found, a carrier can be uniquely identified.

In study III and IV we pooled the matched normal samples ten by ten to increase throughput. As we do not call novel variants in the normal genomes, but rather use it as a negative control to exclude germline SNPs, we do not need to identify the carrier of a variant. To secure the presence of matched normal we used SNPs that were included in dbSNP and called in only one of the leukemic samples.

Identification of driver genes

We assume that if a mutation found in our patients is of importance for the progression of leukemia it would occur in more than one patient. The strongest evidence would be if exactly the same mutation is seen in many cases, but also mutations in the same gene indicate that the gene is of importance. The naïve approach of just counting number of mutations in a gene will be too simple, as it will not take into account the size of a gene and normal frequency of mutations in that genomic region. We have used three programs to identify genes that are significantly mutated in ALL; MutSigCV (37), OncodriveCLUST (38) and OncodriveFM (39). They have all in common that they estimate a normal distribution of mutations from silent (synonymous) somatic SNV and compare it to the distribution of nonsynonymous SNVs (nsSNV). For variants outside the exome the stratification of silent SNVs
and nsSNVs does not exist and it is more complicated to calculate a normal frequency of mutations in the region.
Results

Paper I: Accurate detection of subclonal single nucleotide variants in whole genome amplified and pooled cancer samples using HaloPlex target enrichment

Study I investigated the detectability of somatic single nucleotide variants (SNV) and the conservation of their allele fractions (AF) when using whole genome amplified DNA (wgaDNA), and when mixing non-indexed samples before library preparation.

Validation

Two ALL reference patients were subjected to whole genome sequencing (WGS) of both a leukemic and a matched healthy normal sample. Candidate somatic SNVs were called in any of two programs and filtered against public databases of normal variation and complex regions of the genome, such as repeated regions.

Using Haloplex target enrichment from genomic DNA (gDNA) from both reference patients we were able to validate 532 (227 + 305) SNVs called in WGS. In the WGS data we had called 1541 candidate positions and hence our validation rate was 34.5%. All variants that had been validated with PCR and Sanger sequencing before the targeted sequencing were also validated with Haloplex (n=30). To verify that there was no bias in the selection of the previously validated somatic SNVs a random selection of five SNVs per patients that were validated with Haloplex were also validated with PCR and Sanger and all ten variants were true positives. The set of validated somatic SNVs, together with 19 previously known heterozygote SNPs, were used to evaluate the use of whole genome amplified DNA (wgaDNA) in lieu of genomic DNA and whether it is possible to pool DNA from multiple patients and still detect SNVs. The validation rate might seem to be low compared with other studies (29) but it is important to point out that we performed a genome-wide validation on real data where other groups have focused on validating the exomic variants (40–42). Our validation rate is very similar to what was presented in the original paper describing SomaticSniper (28). SomaticSniper presents a confidence score, SomaticScore, for each variant.
called and Figure 5 illustrates differences in validation rate when using different cut-offs. We used 40 as a cutoff for SomaticScore in all studies, with subsequent validation in high coverage data. In study 1 the candidate variant-calling was performed with both MuTect and SomaticSniper but for the sake of simplicity only scores from one variant caller, SomaticSniper, were used in this illustration. If we raised our cut-off to e.g. a SomaticScore of 70 the validation rate would increase to 78.3% but we would only miss 8.6% of the variants validated at a cut-off of SomaticScore 40.

![Figure 5](image-url)

*Figure 5.* Validation-rate and fraction of true positives lost using different cutoffs on the quality-score from SomaticSniper, “SomaticScore”. The dots represent the fraction of validated candidate SNVs when filtering out all candidates with a lower SomaticScore. The bars represent the fraction of validated SNVs that are lost when filtering on the given SomaticScore compared with keeping all variants with SomaticScore at or above 40, our cutoff before validation in further studies.
Whole genome amplified DNA compared to genomic DNA

SNV-calling on wgaDNA and gDNA from the two reference patients detected 214 and 300 variants in both samples. In gDNA 18 variants were called that were not called in wgaDNA, and 11 variants called in wgaDNA were not found in gDNA. Variants that were inconsistent between gDNA and wgaDNA could be explained by lack of coverage, AF close to cutoff or alignment artifacts rather than problems with the whole genome amplification process. The slightly more uneven coverage of wgaDNA (Study I Figure 2) can explain why wgaDNA had a lower number of validated variants than gDNA. Allele fractions are conserved in the whole genome amplification (Study I Figure 3) and the 19 germline SNPs have AF closer to the expected AF of 0.5 in wgaDNA than in gDNA (Study I Table 2/Figure 3). Our results support the use of wgaDNA as an alternative for detection of SNVs when original material is scarce.

Non-indexed pooling of DNA

We assumed that somatic SNVs would rarely occur at the same position in two different patients or at the same position as a SNP. Therefore the AF of the somatic SNVs are feasible to use as patient-specific tags to determine the representation of the reference patient (ALL_458/ALL_559) in the pool. Using these AFs of the validated somatic variants we calculated an expected AF for each variant in the pool (AF in reference patient divided by the number of patients in the pool). If we compare the expected AF with what we observe in the pool we can quantify how much of the DNA in the pool comes from the reference patient. There is a slight over-representation of the reference patient in all but one pool (Table 2, Study I Figure 4). The observed AF has a higher spread than the expected AF and this seems to correlate with the number of patients in the pool. This is probably due to a lower sequencing depth per patient in the pool as all reactions have approximately the same sequencing depth irrespective of number of samples pooled.

In the pools we detected 97.7-100 % of the somatic SNVs from the reference patients. Undetected SNVs were mostly in regions of low coverage or had a low expected AF. As we presumed somatic SNVs are not present in both patients, variants in pools where the known carrier was not present would then be a false positive. From this we estimated the false discovery rate (FDR) to 0-5.9% (Study I Table 3).

We were also able to identify 6 putative somatic SNVs in any of the 14 additional patients. PCR and Sanger sequencing validated all 6 variants in a unique patient per variant. This SNV-calling, together with the detection of known variants, can be seen as a proof of principle for variant calling in non-indexed pools for further studies.
Table 2. Representation of the whole genome sequenced patient in the pools

<table>
<thead>
<tr>
<th>Pool</th>
<th>Expected1 (%)</th>
<th>Observed2 Original pool (%)</th>
<th>Observed2 Replicate pool (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL_458 + 1 pat.</td>
<td>50</td>
<td>51.6</td>
<td>47</td>
</tr>
<tr>
<td>ALL_559 + 4 pat.</td>
<td>20</td>
<td>24.7</td>
<td>22.6</td>
</tr>
<tr>
<td>ALL_458 + 9 pat.</td>
<td>10</td>
<td>12.1</td>
<td>11.8</td>
</tr>
</tbody>
</table>

1 Representation of one patient in a perfect equimolar pooling (1/number of patients in pool)
2 Median of reference patient-AF/pool-AF for all positions with a somatic SNV

Paper II: Identification of somatic variants by targeted sequencing of pooled leukemia samples

Study II investigated whether it is possible to detect and define the carrier of somatic or rare single nucleotide variants (SNV) in non-indexed pools where samples have been pooled in a redundant scheme. We created two separated pooling schemes based on whether the input-DNA was wgaDNA or gDNA. We then called and attributed variants in 172 patients, where four reference patients (from study III) had a known set of somatic mutations and were used to evaluate the results. We used a region with the exons of 872 genes (same as in study IV) and 907 positions with a known somatic SNV. A matched normal DNA sample was only available for the four reference patients with a known set of somatic mutations.

Verification of pooling

To identify variants from all patients in a pool it is important that they have a uniform representation in the pool. By using the same methodology as in study I we were able to measure the presence of the four reference patients with a known set of variants. As we included 10 patients in each pool, each patient would contribute to 10% of the DNA in the pool under ideal circumstances. We saw that in the wgaDNA-pools all reference patients contributed close to 10% the pool and seemed to be well represented in the pools. For the gDNA pools two patients had a good representation, and two patients had a very low representation. As each patient was included in two pools and the estimated representation of the patient seemed to be consistent between these two pools we argue that representation is more dependent on the measurement of the samples than the manual work distributing the samples in the pools. We redid the pooling of one of the patients with low representation using input-material where the concentration were remeasured and the representation increased from 1.6-2.1% to 7.3-8.1%.
Sensitivity and Precision

Using our four reference patients we were able to evaluate the fraction of known variants called in the pools (sensitivity) and the fraction of calls that are true positives (precision). For patients with a good representation in the pools we see a sensitivity ranging 62.4%-77.4%. A confirmation of the importance of representation in the pools is the patient we remeasured and re-did the pooling, that raised its sensitivity from 9.4% to 59.4%. The precisions were in the span of 92.3%-100% and no correlation with representation in the pools can be seen (Study II Figure 3B). As precision for somatic variants is more related to how well false positives are filtered the lack of correlation is not unexpected.

Novel variants

After filtering and decoding 2146 variants were called at 2041 unique positions. We called on average 12.8 variants per patient range 1-48 variants. From study III and IV we estimate that we call 3-4 more variants in patients with no matched normal compared to patients where we have sequenced a matched normal. It is feasible to assume that about 75% of these variants are rare germline SNPs not found in any public databases. This is in line with the number of variants found in only one individual (singletons) in the 1000 genomes project that is around 3 singletons/Mb and genome in the European population. The region of the genome targeted was roughly 3Mb and on average 9 variants per patients could be expected to be singletons and hence not true somatic events. In study IV we analyzed the same target region and identified 19 putative driver genes. In 14 of those genes we are able to identify a similar distribution of variants in the pools (Study II, Table 3).

Paper III: The mutational landscape in pediatric acute lymphoblastic leukemia deciphered by whole genome sequencing

In this study we investigated the mutational burden on the whole genome of four ALL patients using WGS and RNA sequencing. The patients were selected from four major subtypes of ALL: T-ALL, BCP-ALL of the t(12;21) subtype, BCP-ALL denoted other and BCP-ALL with normal karyotype. We follow up the findings from these four patients in a validation cohort of 168 additional ALL patients to investigate if any of the mutations are recurrent. These 168 patients represent all major ALL subtypes.

From study I we conclude that most variants are detectable in non-indexed pools. Study I was also a pilot for variant-calling where two of the
patients from this study were included. This result was used in the design of study III where we pooled normal samples from patients in the validation cohort before library preparation. Using similar methods as in study I and II we were able to observe the presence of a matched normal for a specific patient in these pools.

Somatic mutations
We found and validated 909 (191-305 per patient) somatic SNVs and exonic indels out of 3088 (713-851 per patient) candidate SNVs, giving us a validation rate of 29.4%. The validation rate might appear to be low compared with other WGS studies of ALL (41,40), but those studies have not presented validated mutations from the whole genome, but instead “high-confidence calls” for regions outside the exons. We detected 15 (2-5 per patient) copy number alterations (CNA). The majority of the CNAs (10/15) were not detected at diagnosis, where a copy-neutral loss of heterozygosity (LOH) on 9p in ALL_559 (T-ALL) and partial deletion of IKZF1 in ALL_501 (normal) could be of most clinical significance of the CNAs undetected with standard protocol at diagnosis (43,44). With RNA-seq we confirmed that all translocations predicted by the cytogenetic analysis result in expressed fusion genes.

Genome-wide mutational patterns
The set of validated SNVs genome-wide enables us to look for patterns in the genome. For two patients (subtypes BCP-ALL with the t(12;21) translocation and T-ALL) > 30% of the mutations seems to be subclonal (Study III Figure 2a) present in 40-60% of the leukemic cells. In patient ALL_458 (t(12;21)) two of the CNAs were subclonal (Study III Figure 2c). As the treatment of these two patients was successful and they are still in remission 5+ years after diagnosis the mutations in the subclone were no additional threat. But for patients that later relapsed this pattern could be of interest to explain why the disease returned.

Patient ALL_707 (subtype other) had an unusual profile of SNV-types, with a majority of C>A mutations. The other three patients had C>T as the most frequent mutation, and this is the most common variant in many other cancers e.g. CLL (34). C>A mutations are correlated with smoking and was found as most frequent in a WGS of a lung cancer (35) but the connection between (parental) smoking and pediatric ALL is not clear and inconsistent reports exist (36).
Recurrent mutations

We designed a panel of 30 genes that had an exonic or UTR mutation in any of the four WGS patients. We also included regions outside the exome based on mutations in regulatory regions. The panel was analyzed in 168 additional ALL patients, where sequences from a matched normal DNA sample were available for 139 patients.

The most frequently mutated genes in the validation cohort were KRAS and NOTCH1, two well-known driver genes both in ALL but also in other cancer-types. A near mutual exclusive pattern between these two genes can be observed where NOTCH1 is only found in T-ALL (11/23) and KRAS is found in 16 BCP-ALL and 1 T-ALL (Study III: Figure 4). After statistical analysis three additional genes were highlighted as potential driver genes; KMT2D, KIF1B and ME1. ME1 were excluded as 2/3 mutations were found in T-ALL, but no T-ALL expressed ME1. Both KMT2D and KIF1B have been linked to cancer, but not ALL, before and are much more likely to be true driver mutations.

Outside the exons we found 12 mutations, but no custom region harbored more than one mutation and it would be hard to distinguish these mutations from random events.

Paper IV: Targeted sequencing of cancer genes in pediatric acute lymphoblastic leukemia

In study IV we expanded the target region from 30 genes in study III to 872 genes, approximately 1‰ of the human genome. The set of genes were based on searches in databases and literature where we found a connection to ALL, cancer or epigenetic regulation. We also expanded the analysis by including 24 relapse samples from 19 of the 172 patients in study III. From these 172 patients 139 with a matched normal DNA sample were considered as a “core cohort”.

We applied the same methods to analyze the data and identify driver genes as for the follow up patients in study III. Using the results from study II we were able to extend the cohort with 168 patients for support of specific mutations.

Somatic mutations in diagnosis and relapse

In our region of interest we detected a total of 1041 somatic SNVs or indels in diagnosis samples. One patient seems to be hypermutated with 120 mutations corresponding to more than 10% of the total. Excluding this patient we found an average of 3.8 mutations per patient in the core cohort and 13.4 mutations per patient with no data from normal DNA. Due to this large dif-
ference only the core cohort has been considered for quantitative differences of mutations between groups. There is a significant difference in number of mutations per patient between T-ALL and BCP-ALL (average of 6.4 and 3.5). There is also a significant difference between these groups in terms of allele fractions of the mutations where T-ALL patients have a significantly higher proportion of mutations stemming from the founder clone.

In relapse samples we detected in total 94 mutations not detected in their matched diagnosis sample. A distinctive characteristic of new mutations in the relapse samples is that they seem to be accumulated to genes selected due to their involvement in epigenetic regulation (Study IV, figure 1D).

Driver genes

Screening for putative driver genes was performed as in Study III. Patients were divided into five subsets based on subtypes before driver gene analysis and 19 genes were identified as putative driver genes in any of our subsets (Study IV, figure 2). Four of these are part of the Ras signaling pathway, affecting predominantly patients with BCP-ALL. Five of the putative driver genes (DNAH5, ATRX, FUBP1, ABCB5 and SYNE1) have not previously been linked with ALL and all are predicted to be drivers only in BCP-ALL (Study IV, Figure S3).

Survival analysis

Using Gray’s test (45) we observed a significant association between mutations at diagnosis in the Notch signaling pathway and relapse BCP-ALL patients. Of the genes included in this pathway only CREBBP was identified as a putative driver gene on its own. This gene has been linked to relapse earlier (46) but detailed analysis of our patients suggests that the association to relapse is not driven only by CREBBP. No other pathway or single gene mutations at diagnosis were significantly associated to relapse in our dataset. The genes in the Notch signaling pathway were further analyzed in our set of relapse samples. We observed that two patients gained a mutation in EP300 at relapse. In the extended dataset from study II one patient that later relapsed had a mutation in EP300 at diagnosis. This observation together with the absence of EP300 mutations in patients that have not relapsed indicates further that this gene might be associated with relapse.
Summarized discussion

This thesis describes methodological development of a sequencing experiment, and application of these methods in large projects.

In study I we showed that using whole genome amplified DNA (wgaDNA) is an alternative if the amount of original DNA is scarce. In Study II-IV we could increase the size of the cohort by more than 10% using wgaDNA. Seemingly small it might have an impact in studies of leukemia and other types of cancer where samples are few. In retrospective studies the available material might be scarce and whole genome amplification is the only way to investigate valuable samples. As we only investigated SNVs we did not evaluate other artefacts linked to wgaDNA, e.g. overcalling of inversions (47). We also presented a proof of concept for variant calling of putatively somatic variants in samples pooled without prior indexing of individual samples.

In study II we evaluated the possibilities to detect mutations in non-indexed pools and assign them to carrier using a known set of variants from four reference patients. Earlier studies have used similar strategies to detect rare germline variants (36,48) but to my knowledge we are the first using it to detect somatic variants. We were also able to quantify the representation of the reference patients in the pools. Using this pooling scheme together with the strategy of pooling matched normal DNA samples, as done in study III and IV, it is possible to increase the number of patients sequenced per targeted capture six times1. This, however, does not take into account the cost of sequencing and the method should only be applied on projects where few genes are captured and the cost of sequencing is low in comparison to fixed costs (e.g. library preparation). A challenge in study II was the absence of constitutional matched normal DNA samples, which complicated the search for somatic mutations. To overcome this obstacle one might prioritize functional annotations as in (49). The establishment of a Swedish reference database for normal variation will also be useful due to the lack of Swedish samples in large sequencing projects such as 1000 genomes (50).

1 In the pooling scheme we use 10% of 2 reactions with leukemic DNA and 10% of 1 reaction with matched normal samples. In total 0.3 reactions; If no pooling is performed we use 1 full reaction for leukemic DNA and 1 full reaction for matched normal DNA sample. In total 2 reactions
**Study III** explored the genomic landscape of four ALL patients. Our genome-wide validation enabled detection of patient-specific patterns of mutation types. Although it is hard to determine the importance of mutations outside the exome they might be helpful to reveal the underlying causes of cancer. We observed an unusual pattern of mutations normally associated with smoking. On a larger scale this pattern has been observed in lung cancer (51). We successfully identified subclones of cells in two patients from the distribution of allele fractions of the SNVs. At the time of publication for study III subclones in ALL had been described for copy number alterations (52) or assumed for few exonic mutations (53) but not at a genome-wide level.

In **study IV** we screened 872 genes to find recurrently mutated genes in ALL and to what extent they can be associated with relapse. The cohort screened includes all major subtypes of ALL. We identified two novel putative driver genes (SYNE1 and ATRX). These two genes have been identified as recurrent in different types of cancer, although never in ALL.

Patients with BCP-ALL were found to have a higher rate of relapse if they carried a mutation in the notch signaling pathway. This pathway contains CREBBP that has earlier been linked to relapse in BCP-ALL (46), although patients in our cohort with mutations in that gene do not relapse in a higher rate than patients with a mutation in other genes of the notch signaling pathway. Another gene of interest in the notch signaling pathway is EP300. One patient had a non-synonymous mutation in this gene at diagnosis, two other patients at relapse and one patient at diagnosis in our extended cohort from **study II**. These four patients were all BCP-ALL from four different subtypes and all of them experienced a relapse. Mutations in EP300 have only been observed as singular events occurring at relapse in earlier studies (46,54) and would not have been detected by us if we had focused only on specific subtypes. For one of the patients that acquired a EP300 mutation at relapse it was possible to detect cells with the mutation at diagnosis. Subclones of leukemic cells that resist treatment have been observed earlier in ALL (55–57) and it could be of clinical importance if a targeted therapy against EP300 were available. Clinical trials with Histone deacetylase inhibitors are ongoing in different types of lymphomas and could be a solution for ALL as well (58).
Acknowledgement

The work described in this thesis was performed in the Molecular Medicine group at the Department for medical sciences, Uppsala. Numerous of people have contributed to this thesis. My special appreciation goes to:

My main supervisor Ann-Christine Syvänen: Thank you for giving me the opportunity to do my PhD in your group. Your enthusiasm for the project, experience in the field and ability to see “the next thing” have been an inspiration for me. I appreciate that you never hesitate to give your honest opinion on any issue and always support new ideas.

My co-supervisor Eva Berglund. Without you this thesis would not exist at all! For always being there to discuss ideas and projects. Thank you for your great patience when I were pushing deadlines. For your enthusiasm and always encouraging me. For all the help with writing.

All clinicians involved in the treatment of ALL patients: Thank you for good collaboration. Especially Gudmar Löönerholm for all your work with ALL in Uppsala, enabling this thesis and Erik Forestier for introducing me to cytogenetics. ISCN has become an important resource to decode the karyotyping strings.

Past and present members of the Molecular Medicine group: Elin, Amanda, Niklas, Sara, Sara and Ying, Thank you for doing the labwork included in this thesis. Behrooz, Shahina and Satyam, Thank you for helping out with the analyses. All junior PhD-students: Johan, for all help with alignment and scripting, Juliana and Yanara for taking over the MDR/PhD-rep-baton. And Juliana, you are senior soon! 😊 Christofer, Anders and Jonas thank you for all R and Python assistance. Tom, Juliana and Nour for proof reading parts of the thesis. Per for sharing your knowledge of genetics and help with relocation. Lillebil and Katarina for helping out with all administrative tasks. Jessica, Anna, Chuan, Johanna, Annika, Josefin, Lina and Sofia and all former PhD students, lab-assistants and Postdocs in the group. Jessica Nordlund, Thank you for sharing all your knowledge of ALL and inspiration for the pooling ideas. Michelle, Erika, Mathias and Katerina for providing a nice atmosphere in the group.
Everyone at the SNP&SEQ platform. As the project has been so large probably everyone at the platform have been involved with the data in this thesis at some point. Thank you for your excellent work providing high quality sequencing data.

Diana Ekman and Anna Johansson at Bioinformatics Long-term Support (WABI) for help with setting up pipelines and Eva Freyhult at Bioinformatics Short-term Support and Infrastructure (BILS) for statistical assistance.

Everyone at CKMF, especially Anna-Stina, Jessica and Katarina who introduced me to PhD-studies, made me book a dissertation date well in advance and accepted my noisy espresso machine.

All PhD reps at MedSci: Oskar, Christofer, Nils, Henning and Markus for good times, pubs and PhD-day preparations. Oskar, thank you for showing the right doors to knock and good luck in Boston! Colleagues in the MDR board: Johan, Camilla, Carl-Johan and Andreas. Pontus, Mareike, Daniel and all other PhD students in Uppsala.


Lisbeth, Jan, Magnus and Victoria: for helping us out during this intense period and taking excellent care of Calle and Erik. Lisbeth for your wonderful food packages which have saved us during hectic days. Jan for inspirational conversations about the thesis and thank you for letting me use your “writing cottage”.

My parents and brother for always being there. Dad, Thank you for teaching me to keep calm no matter the situation and for inspiring me to study mathematics. Mom, Thank you for teaching me to be curious and taking excellent care of Calle and Erik both in Norje and Västerås.

My extra-parents Marita “Minna” and Stor-Anders Rystedt. Thank you for your 24/7 support with food, childcare, housing, scheduling, proof reading etc in the process of writing this thesis.

Mina tre solstrålar: Ida, Carl och Erik. Tack för att ni är ni!
References


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1170

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine”.)