Multi-loci diagnosis of acute lymphoblastic leukaemia with high-throughput sequencing and bioinformatics analysis

Yann Ferret, Aurélie Caillault, Shéhérazade Sebda, Marc Duez, Nathalie Grardel, Nicolas Duployez, Céline Villenet, Martin Figeac, Claude Preudhomme, Mikaël Salson, et al.

Multi-loci Diagnosis of Acute Lymphoblastic Leukaemia with High-Throughput Sequencing and Bioinformatics Analysis

Yann Ferret* Aurélie Caillault* Shéhérazade Sebda Marc Duez
Nathalie Grardel Nicolas Duployez Céline Villenet Martin Figeac
Claude Preudhomme Mikaël Salson* Mathieu Giraud*

Final accepted manuscript, as published in ***British Journal of Haematology, 173:413-420 (2017)***
http://dx.doi.org/10.1111/bjh.13981

Summary

High-throughput sequencing (HTS) is considered a technical revolution that improves our knowledge of lymphoid and autoimmune diseases, changing our approach of leukaemia both at diagnosis and during follow-up. As part of an Ig/TCR based minimal residual disease (MRD) assessment of acute lymphoblastic leukaemia patients, we assessed the performance and feasibility of the replacement of the first steps of the approach based on DNA isolation and Sanger sequencing, by a high-throughput sequencing protocol combined with a bioinformatics analysis and a visualisation using the Vidjil software. We prospectively analysed the diagnostic and relapse samples of 34 paediatric patients, thus identifying 125 leukaemic clones with recombinations on multiple loci (TRG, TRD, IGH and IGK), including Dd2/Dd3 and Intron/KDE rearrangements. Sequencing failures were halved (14% vs 34%, $p = 0.0007$), enabling a monitoring of more patients. Furthermore, more markers per patient could be monitored, reducing the probability of false negative MRD results. The whole analysis from the sample reception to the clinical validation was shorter than our current diagnostic protocol, with equal resources. The V(D)J assignation was successfully done by the software, even on unusual recombinations. This study emphasises the progress that HTS with adapted bioinformatics tools can bring to the diagnosis of leukaemia patients.

Introduction

Molecular Diagnosis of ALL

In Acute Lymphoblastic Leukaemia (ALL), Minimal Residual Disease (MRD) negativity is currently recognised, at least for Philadelphia-negative ALL, as the best independent prognostic factor (Brüggemann et al. 2012). The correct quantification of MRD is used for the initial assessment of treatment response, risk stratification, and later for early relapse detection (Cavé et al. 1998; Gandemer et al. 2014). Several techniques can be used to accurately quantify MRD, including Ig/TCR MRD, which is the most standardised approach throughout Europe (van Dongen et al. 2002, 2003). It is based on a step by step analysis of the V(D)J DNA rearrangements within the lymphoblast. First, leukaemic clonal rearrangements are amplified by PCR and searched by capillary electrophoresis, they are then isolated using a polyacrylamide gel electrophoresis and are finally sequenced using a Sanger based approach. For each marker (ideally 2 per patient), an allele-specific oligonucleotide (ASO) is designed in order to quantify MRD by real time Q-PCR (TaqMan chemistry).

It is recommended to follow at least two markers per patient with ASO real time Q-PCR. Indeed, false negative early MRD measurements sometimes occur when the followed marker(s) is (are) specific of a chemo sensitive clone which is accompanied by a non-followed chemo resistant clone. ALL is an oligo clonal disease that undergoes a certain degree of clonal evolution from diagnosis to relapse. Relapse can be due to unmonitored abundant clones as well as sub-clones or new clones, causing false negative late MRD measurements (Kerst et al. 2005). The quality of the MRD measurement is therefore closely related to the number of followed markers. In 95% of the cases a potential marker is found. However, for a given sequence, it is not always possible to design an allele-specific oligonucleotide which is capable of providing the required analytical sensitivity. 90% of patients can be effectively monitored. The risk of false negative MRD results – as described above – increases when only a single rearrangement is monitored. It is therefore important to obtain several markers for each patient.

Repertoire Sequencing (Rep-Seq) for ALL diagnosis and MRD

Studies have assessed the usefulness of high-throughput sequencing (HTS) for repertoire sequencing (Boyd et al. 2009; Logan et al. 2011; Gawad et al. 2012; Faham et al. 2012; Logan et al. 2013, 2014; Wu et al. 2012). A Repertoire Sequencing (Rep-Seq) study consists of the deep sequencing of a lymphoid population, focusing on V(D)J recombinations (Tonegawa 1983; Market and Papavasiliou 2003). Such studies have been used to identify repertoires in animals (Weinstein et al. 2009; Ben-Hamo and Efroni 2011; Castro et al. 2013) and humans (Freeman et al. 2009; Warren et al. 2009; Boyd et al. 2010; Warren et al. 2011; Arnaout et al. 2011; Weng et al. 2013; Laserson et al. 2014; Niklas et al. 2014).

In ALL, HTS at diagnosis may reduce the number of sequencing failures: While it is sometimes difficult to isolate a clone on polyacrylamide gel, HTS does not suffer from such pitfalls. It is therefore possible to accurately monitor more patients, with more markers, and a technology which is both faster and easier to use.

During MRD follow-up, HTS may provide a global, thorough, deep and forever-evolving vision of the aberrant lymphoid populations thus giving far more information with a lighter process than ASO Q-PCR.

Challenges

Given that the purpose is to identify the major leukaemic clones at diagnosis, a few dozens of thousands of reads over the V(D)J junctions for each locus are enough. One needs a reliable and reproducible method. Replacing ASO Q-PCR is a more intricate point to achieve because it requires higher accuracy and lower analytical sensitivity for every possible V(D)J rearrangement studied. That would require a larger amount of DNA as well as correction using spikes of known concentration (van Dongen et al. 2015). This is not the aim of this work.

In both cases, the huge amount of data raises two challenges. First, diagnostic laboratories must be able to store and process terabytes of data per year. Secondly, the data must be nicely synthesised to ease clinician interpretation. Moreover, there is a strong need for standardisation by an established consortium, such as, in Europe, the ESLHO, in order to obtain comparable results within the different centres.

Repertoire Sequencing (Rep-Seq) analysis software

With the rise of high-throughput sequencing, there is a need for dedicated software able to analyse high-throughput data containing CDR3 (Complementary determining Region 3) (Benichou et al. 2012). Indeed, the analysis of V(D)J recombinations poses a very specific problem which cannot be solved with generic error-correction, mapping or clustering tools in an acceptable way. Algorithmic methods should take into account the V(D)J recombinations to appropriately handle small recombinations, somatic hypermutations, or short insertions.

The international ImMunoGeneTics information system (IMGT®) has developed several tools for the in-depth analysis of V(D)J recombinations (Yousfi Monod et al. 2004; Brochet et al. 2008; Lefranc 2011; Alamyar et al. 2012)[1]. Recently, new software able to deal with up to millions of sequences has been designed: (Arnaout et al. 2011), IgBlast (Ye et al. 2013), Decombinator (Thomas et al. 2013), miTCR (Bolotin et al. 2013), Vidjil (Giraud et al. 2014), TCRKlass (Yang et al. 2014), MiXCR (Bolotin et al. 2015), IMSEQ (Kuchenbecker et al. 2015). At the heart of these programs is the optimised comparison of the reads against germline databases, and, for some of them, clusterization of the reads into clones.

Contents

In this paper we report on our experience of a 3 months long prospective study of diagnostic paediatric ALL samples in a routine hospital practice. As part of the Ig/TCR MRD follow-up, we assessed relevance, performance and feasibility of the replacement of the current protocol (isolation and Sanger sequencing of leukaemic markers), by a HTS protocol combined with a bioinformatics analysis and an interactive visualisation with the new Vidjil software (Giraud et al. 2014).

[1] `http://imgt.org/`

We studied recombinations on 36 samples for multiple loci (TRG, TRD, IGH and IGK), including Dd2/Dd3 and Intron/KDE (Kappa Deleting Element) rearrangements.

Patients and methods

Patient selection

From January to March 2015 we prospectively included every paediatric patient (0 to 25 years old accepted) newly diagnosed with an ALL, or who relapsed during this period, that were hospitalised in the CHRU of Lille or in a partner hospital (CHU of Lyon, Rouen, Amiens, Poitiers, Reims and CH of Marseille La Timone), and had an available bone marrow sample at diagnosis or at relapse time (at least 80% blasts). The approval for this study was obtained from the Institutional Review Board of CHRU of Lille (CSTMT093) and was in accordance with the Declaration of Helsinki regarding the informed consent of patients. For each patient, a written informed consent was obtained from the parents or legal guardians, or from the patients themselves when they were older than 18.

34 paediatric patients were included: 32 with a diagnosis sample, 2 with both a diagnosis sample and a relapse sample, totalling 36 samples. The current Ig/TCR technique and HTS technique were implemented separately by two different teams. The results were also analysed separately by different people before comparison.

DNA extraction and amplification

For every sample, genomic DNA from 5M cells was extracted on bone marrow with QIAamp® DNA Mini Kit (Qiagen). DNA concentration was measured on Nanodrop system® and 125ng were amplified by 5 multiplexed PCR systems with a high fidelity polymerase (FastStart^{TM} High Fidelity PCR System dNTPack, Roche). These systems are described or derived from the Biomed-2 group works (see Table 1). Most loci of TRG, TRD, IGH and IGK are thus amplified. Amplicons were checked on 1,5% agarose gel, then purified on AMPure® beads and quantified with Quant-iT^{TM} PicoGreen® dsDNA kit.

Preparation of libraries

The 5 PCR products were pooled in an equimolar amount and then sequenced with a unique barcode. The libraries were prepared using the Ion Plus Fragment Library Kit (01/31/2012) (Life Technologies). Each library was barcoded using the Ion Xpress^{TM} Barcode Adapters Kit. The concentration of each barcoded library was controlled using the 2200 TapeStation system (Agilent Technologies), then they were pooled and their concentration was checked with the Agilent 2100 Bioanalyzer.

Emulsion PCR, enrichment and High-throughput sequencing

The amplification of the libraries by emulsion PCR was performed using the Ion OneTouch^{TM} 2 (Life Technologies) and Ion PGM^{TM} Template OT2 400 Kit (Life Technologies). After amplification the libraries were enriched using the Ion OneTouch^{TM} ES (Life Technologies). The estimate of the proper conduct of the PCR was audited with the Ion Sphere^{TM} Quality Control Kit on the Qubit® 2.0 Fluorometer (Invitrogen). The libraries were sequenced on Ion Torrent Personal Genome Machine (Ion PGM^{TM}) using Ion PGM^{TM} Sequencing 400 Kit and Ion 316^{TM} Chip Kit (Life Technologies, 8 samples per chip).

Bioinformatics analysis and visualisation

The raw Ion Torrent flow was transformed to demultiplexed sequences with the Torrent Server from Life Technologies allowing 0 errors in barcode splitting.

The reads were processed by the Vidjil software (www.vidjil.org, versions 2015.04 and 2015.05, with default parameters). Vidjil gathers the reads into clones on the basis of their V(D)J rearrangements. Two reads belong to the same clone if they share a same 50 bp "window" overlapping the actual CDR3. This window is large enough to be highly specific. The window detection is based on a fast bioinformatics method using k-mer indexing of the germline genes (Giraud et al. 2014).

We explored the lymphocyte populations on the multiple loci of the 36 samples in the Vidjil browser, tagging 125 recombinations as clones of interest. These recombinations are shown in the Supplementary Table 1. The interactive visualisation of these populations can be accessed at `www.vidjil.org/bjh-2016`.

Because of the sequencing errors, there may be several bioinformatics clones corresponding to a real clone. In such cases, the user can choose to merge them in the browser. We did so when the pattern of error clearly indicated a sequencing error (see Figure 2). Vidjil then outputs a representative sequence for each of the most represented clones Every segment of the representative sequence is guaranteed to be shared by at least 50% of the reads assigned to the clone (Giraud et al. 2014). The V(D)J recombination analysis (determination of V(D)J germline genes and N region) is done only at the end of the pipeline, using either Vidjil built-in analysis, IMGT/V-QUEST (Brochet et al. 2008), or IgBlast (Ye et al. 2013).

Results

36 samples from 34 patients were included (28 B-ALL / 6 T-ALL, 17w / 17m, 0-25 years). There was between 77,041 and 1,157,040 reads per sample (median 338,043). The median ratio of reads recognised as V(D)J rearrangements by Vidjil was 86.8%, giving a median number of reads of 262,448 per patient.

Identification of the main clones

The intention is to compare HTS to the conventional capillary electrophoresis technique. In the 36 samples, 145 clones were identified by capillary electrophoresis versus 125 by HTS. 11 of the 20 missing clones are the predictable consequence of the use of a primer in the IGH system called "JHDN" which has a weak ability to amplify rearrangements involving *IGHJ3* or *IGHJ6* genes. A subsequent experiment with two additional JH primers lead to the identification of 10 out of these 11 missing clones, the last sample having an insufficient amount of DNA. The other 9 missing clones were related to insufficient amplification by the initial PCR. Their amplicons were excluded and were not loaded on the HTS chip: 3 in the same sample because of poor DNA quality, 4 related to an ageing primer aliquote, and 2 with very small electrophoretic peaks.

The lengths of the consensus sequences reported by HTS are very close to the lengths of the fluorescent fragments measured on capillary electrophoresis below 300 bp (less than 3 bp mean length difference for TRG, TRD and IGK, see Table 2). For IGH, this difference is around 40 bp. This may be explained by the sequencing protocol used on the Ion Torrent (850 incorporations). These reads are still very long on both sides of the CDR3, neither impeding analyse nor design of the ASOs.

Sequencing

The procedure of isolation on polyacrylamide gel and Sanger sequencing was conducted for 107 leukaemic clones among the 145. The failure rate was 36/107 (34%) whereas it was only 15/107 (14%) with HTS. Successful sequencing was therefore significantly more frequent with HTS ($p = 0.0007$, bilateral Mc Nemar test). In three patients the MRD follow-up was impossible without HTS: There were respectively 6, 2 and 2 visible clonal rearrangements in capillary electrophoresis, but Sanger sequencing failed for every clonal rearrangement. The HTS successfully sequenced 3, 2 and 2 of these rearrangements, enabling the follow-up of the MRD by ASO Q-PCR in these children.

64 clones were successfully sequenced by both approaches. We compared them for more than 50 bp before and 50 bp after the junctions. No significant differences were found.

Reproducibility of PCR and HTS

On the diagnostic sample of another paediatric patient, we performed two types of replicates. We did the initial PCR twice (called PCR A and PCR B). Then for a given PCR, half of the amplicons were labelled with a barcode, and the other half with another barcode. This allows the identification of potential biases during the library preparation and the sequencing. We are therefore able to analyse both the reproducibility of the technique and the steps within the protocol that are less reproducible. This is shown in Figure 1. Some points are negative in one PCR but positive at 5×10^{-4} in the other one. Such extreme cases are not apparent for the sequencing. However, even across different PCR, the main clones are still very well identified: for all clones detected above 10^{-3}, the difference between both PCR is one half-log or less. At lower concentrations, the spread of the distribution could come both from PCR or HTS bias but also from biological sampling. This is not a problem for the diagnosis as we are not interested in very accurate quantification but only in identifying the major clones.

Handling sequencing errors

The sequencing errors depend on the sequencer used (Ross et al. 2013). On the Ion Torrent technology, the most frequent errors are homopolymers. Figure 2 displays such a case on a Vk-Kde recombination. These sequences can be manually or automatically clustered on the Vidjil browser. This is left to the expertise of the

TRG 1-10	Vgf1	+	5′ GGA AGG CCC CAC AGC RTC TT 3′
(BIOMED-2)	Vg10	+	5′ AGC ATG GGT AAG ACA AGC AA 3′
	J1J2	−	5′ GTG TTG TTC CAC TGC CAA AGA G 3′
	JP1/2	−	5′ TTA CCA GGC GAA GTT ACT ATG AGC 3′
TRG 9	Vg9	+	5′ CGG CAC TGT CAG AAA GGA ATC 3′
(BIOMED-2)	J1J2	−	5′ GTG TTG TTC CAC TGC CAA AGA G 3′
	JP1/2	−	5′ TTA CCA GGC GAA GTT ACT ATG AGC 3′
TRD	Vd1	+	5′ ATG CAA AAA GTG GTC GCT ATT 3′
	Vd2	+	5′ ATA CCG AGA AAA GGA CAT CTA TG 3′
	Dd2	+	5′ AGC GGG TGG TGA TGG CAA AGT 3′
	Dd3	−	5′ TGG GAC CCA GGG TGA GGA TAT 3′
	Jd1	−	5′ GTT CCA CAG TCA CAC GGG TTC 3′
	Ja29	−	5′ GGC AAA AGC ATT CTA GGT ACA 3′
IGH	FR1	+	5′ AGG TGC AGC TGS WGS AGT CDG G 3′
	JHDN	−	5′ ACC TGA GGA GAC GGT GAC CAG GGT 3′
IGK	VK1f6	+	5′ TCA AGG TTC AGC GGC AGT GGA TCT G 3′
(BIOMED-2)	VK2f	+	5′ GGC CTC CAT CTC CTG CAG GTC TAG TC 3′
	VK3f	+	5′ CCC AGG CTC CTC ATC TAT GAT GCA TCC 3′
	VK4	+	5′ CAA CTG CAA GTC CAG CCA GAG TGT TTT 3′
	VK5	+	5′ CCT GCA AAG CCA GCC AAG ACA TTG AT 3′
	VK7	+	5′ GAC CGA TTT CAC CCT CAC AAT TAA TCC 3′
	Intron	+	5′ CGT GGC ACC GCG AGC TGT AGA C 3′
	KDE	−	5′ CCT CAG AGG TCA GAG CAG GTT GTC CTA 3′

Table 1: Forward (+) and reverse (–) primers used in the 5 PCRs.

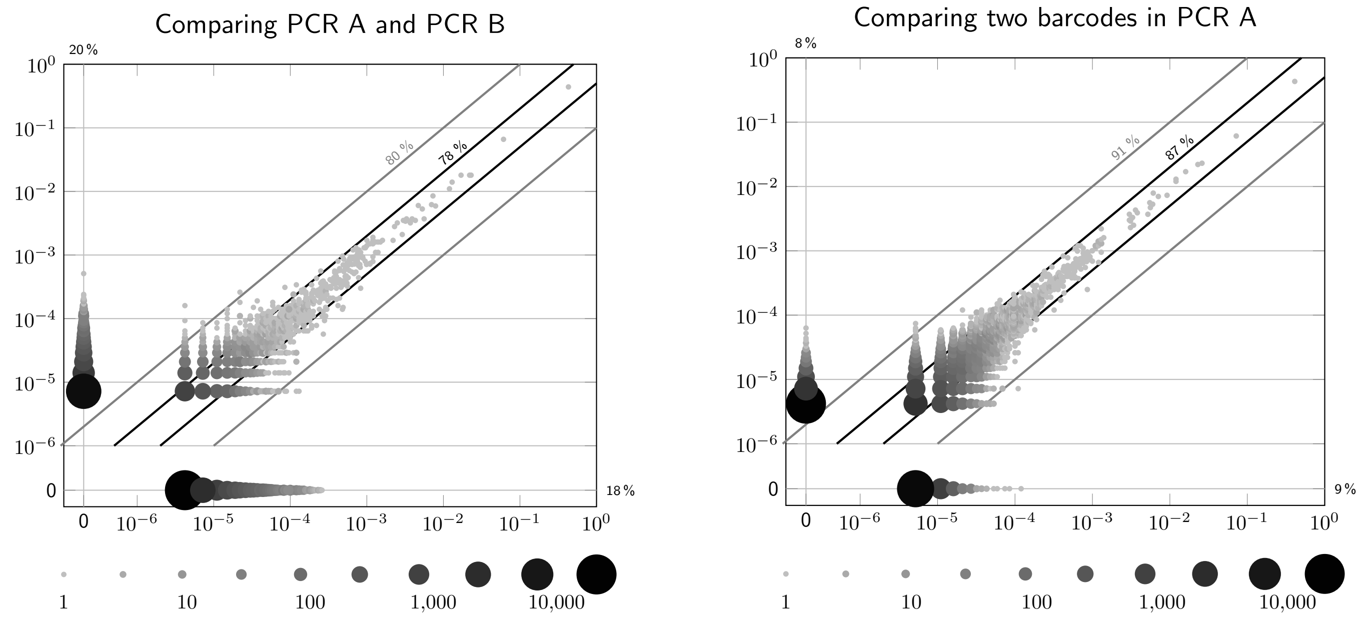

Figure 1: Reproducibility of the PCR (left) and of the sequencing (right) for patient 0014 at diagnosis. Each point represents one "window" as identified by Vidjil (that includes the CDR3 sequence). The ratios are computed on the number of reads, without additional clustering. Since many points may share the same ratios on both axes, the colour represents the number of points having the same ratios: The darker the colour, the more points share the same ratios.

	TRG		TRD		IGH		IGK		total
	n	ℓ	n	ℓ	n	ℓ	n	ℓ	n
Capillary Electrophoresis	54	205.6	41	245.4	31$\star$	373.5	19	285.3	145
HTS	53	205.3	37	248.0	18	330.7	17	272.6	125
fragment length difference		-0.13 ± 14		2.11 ± 4.53		-40.61 ± 35.62		-2.53 ± 15.99	

Table 2: Number (n) and average length (ℓ) of clones detected in each of the four studied locus. Note that the JH6 and JH3 primers were included in the standard technique but not in the HTS protocol.

```
clone-001: window                   AATAA-CTGGCCTCATTTTCTCCC-TATGGGCCTAGTGGCAG
>IGKV3-15*01            ...CAGCAGTATAATAA-CTGGCCTCc
>KDE                                                          ggagcCCTAGTGGCAGCCCAGG...
clone-001   59523  15.3%   CAGCAGTATAATAA-CTGGCCTCATTTTCTCCC-TATGGGCCTAGTGGCAGCCCAGG
clone-008    1766  .454%   CAGCAGTATAATAAACTGGCCTCATTTTCTCCC-TATGGGCCTAGTGGCAGCCCAGG
clone-021     648  .167%   CAGCAGTATAATAA-CTGGCCTCATTTTCTCCCCTATGGGCCTAGTGGCAGCCCAGG
clone-030     540  .139%   CAGCAGTATAATAA-CTGGCCTCATTT-CTCCC-TATGGGCCTAGTGGCAGCCCAGG
                                         h            h     h
```

Figure 2: Representative sequences of the largest clones identified as IGKV3-15*01 -1/16/-5 KDE in the diagnosis of patient 0074, aligned with IMGT/GENE-DB relevant germline gene and KDE (Genbank NG_000834). For each clone, the number of reads and the proportion among all analyzed reads are given. The three positions, marked by `h`, are likely sequencing errors in homopolymers.

user to distinguish sequencing or PCR errors from somatic mutations as it depends on the sequencing technology as well as the PCR.

On our dataset, a manual clustering gathered an average of 8.2 sequences per clone cluster. It did not significantly change the quantification of these clones: The mean size of the main sequence of the clone cluster is 83% of the total size of the cluster.

Comparison between Vidjil, IMGT/V-QUEST and IgBlast on V(D)J designation

We checked the V(D)J designation returned by Vidjil against other popular tools. The consensus sequences of the 125 clones were sent to IMGT/V-QUEST (version 3.3.2) (Brochet et al. 2008) and IgBlast (version 1.4.0) (Ye et al. 2013). The obtained V(D)J designations were manually checked and compared to the result of Vidjil (version 2015.04 and 2015.05).

In 63% of the cases, the three tools perfectly match or give slightly different designations that are acceptable (Table 3). In the other cases, at least one of the programs does not give an acceptable answer. This mostly comes from unusual rearrangements (*TRDD3*, KDE, *TRAJ29*) that are handled neither by IMGT/V-QUEST, nor by IgBlast. Vidjil correctly recognises these recombinations. Most Vidjil errors are due to slightly imprecise junction positioning. Globally, Vidjil gave a correct analysis in 113 out of 125 clones.

Concordance of the three tools			
Same designation		58	(46)%
Minor differences		21	(17)%
Significant differences		46	(37)%
IMGT/V-QUEST and IgBlast			
Correct designation (at least one tool)		77	(62)%
No designation or bad designation		48	(38)%
TRDD3	29		
TRDD3-TRAJ29	1		
TRAJ29	1		
KDE	17		
Vidjil			
Correct designation		113	(90)%
No designation or bad designation		12	(10)%
Bad detection of central gene	2		
TRDJ2 instead of *TRAJ29*	2		
Bad junction detection	7		
No designation	1		

Table 3: Comparison between V(D)J designations made by IMGT/V-QUEST (version 3.3.2), IgBlast (version 1.4.0) and Vidjil (versions 2015.04 and 2015.05) on the 125 clones identified by HTS, and cases of failure of the different tools.

Discussion

Quantity of DNA

The quantity of DNA used (125ng) is sufficient to identify the major clones for diagnostic samples. The quantity to be used for MRD should be carefully studied to ensure both a good representativity of the population van Dongen et al. (2015) and a good PCR efficiency with recent high-fidelity enzymes.

Sequencing quality

Sequencing failures are halved with our HTS method. Therefore we could monitor MRD in more patients. We could also respect the recommendation of 2 markers followed per patient more often. This reduces the risk of false negative results. We believe we can improve IGH locus analysis by replacing JHDN primer with a selection of primers amplifying all IGHJ genes.

Fragment length

The fact that sequence length is usually close to fragment length in capillary electrophoresis (below 300 bp) can be a useful tool for data analysis. For IGH (over 300 bp), the difference in fragment length is coupled with a negative quantitative bias of this whole system over the others. This bias does not prevent from analysing and clustering IGH CDR3. No intra-system bias was observed.

V(D)J designation

Vidjil gave a correct analysis in 113 out of 125 clones, and was the best tool among the ones we evaluated for this purpose, especially on Dd3 and KDE rearrangements. Incorrect V(D)J designation constitutes a major loss of analysis time at the ASO design moment, as manual designation is a complex and error-prone task.

Reproducibility and contaminations

The limiting factor of reproducibility is PCR, using a high-fidelity polymerase is therefore essential. It also highlights the high reproducibility of sequencing. Repeated manipulation of amplicons before ligation to a barcode sequence induces a risk of sample contamination. We did not encounter such unwanted events in our work. Nowadays this risk can be reduced (by using primers which already include a specific barcode sequence for example). However because of the high number of primers involved for each patient, they could not be used in this setting.

Bioinformatics analysis and visualisation

The Vidjil software facilitates the analysis of reads by grouping them in clones, and performs the analysis of both diagnostic and follow-up samples even when multiple loci are studied. Moreover it can be easily and conveniently operated by a clinician, with a web application to upload sequences, to launch the analysis process on a server, and then display and export the results. Vidjil is open-source and yields reproducible and traceable results. Clinicians maintain control over patient data.

Feasibility in haematological lab

The proposed protocol has been designed and tested retrospectively for two years in Lille hospital. It is now used prospectively in Lille and is currently being tested in other French diagnostic labs involved in MRD testing. The whole analysis for a set of samples takes four to five days, including sample preparation, amplification, sequencing, analysis and validation which turns out to be faster than the current standardised Ig/TCR protocols. We achieved the required standardisation level to now make it feasible in routine haematology laboratories. The running costs for this protocol are relatively low and depend on the number of samples sequenced on a chip. The dominant cost is still the acquisition of the new equipment. Moreover, this new multi-step procedure requires experienced workers with high-throughput sequencing.

Conclusion and perspectives

HTS coupled with a bioinformatics analysis and visualisation can nowadays be used in a routine hospital practice. It represents a major step forward in ALL Ig/TCR at diagnosis. In the long run, the aim is to use HTS for MRD follow-up, to replace ASO Q-PCR, which is time consuming, labour intensive and associated with a high false negative rate. It gives a more complete view of the lymphoid population at diagnosis. Furthermore it can then be used to observe the qualitative and quantitative evolution of this whole population. Moreover, it prefigures new steps not only for MRD monitoring in ALL but also for the understanding of other clonal lymphoid proliferations, and immune responses that may be observed in infectious or auto-immune contexts.

Acknowledgements

This research was supported by SIRIC ONCOLille, Grant INCa-DGOS-Inserm 6041. The authors are grateful to Eileen Boyle and Ryan Herbert for their comments on the manuscript.

Competing interests

The authors have no competing interests.

Author contribution

YF, AC, MS and MG are equal contributors. YF, AC, NG and CP selected the patients. MF, SS, CV and AC designed the sequencing protocol. MD, MG and MS conceived the Vidjil software. MS, MG, AC, NG, MF, SS and YF analysed and discussed the data. YF, MS and MG drafted the paper. All authors corrected the paper and approved the final manuscript.

References

Alamyar, E., Giudicelli, V., Li, S., Duroux, P., & Lefranc, M.-P., (2012). IMGT/HighV-QUEST: the IMGT® web portal for immunoglobulin (IG) or antibody and t cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research*, **8** (1).

Arnaout, R., Lee, W., Cahill, P., Honan, T., Sparrow, T., Weiand, M., Nusbaum, C., Rajewsky, K., & Koralov, S. B., (2011). High-resolution description of antibody heavy-chain repertoires in humans. *PLoS ONE*, **6** (8):e22365.

Ben-Hamo, R. & Efroni, S., (2011). The whole-organism heavy chain B cell repertoire from zebrafish self-organizes into distinct network features. *BMC Systems Biology*, **5** (1):27.

Benichou, J., Ben-Hamo, R., Louzoun, Y., & Efroni, S., (2012). Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology*, **135** (3):183–91.

Bolotin, D. A., Shugay, M., Mamedov, I. Z., Ekaterina V Putintseva, M. A. T., Zvyagin, I. V., Britanova, O. V., & Chudakov, D. M., (2013). MiTCR: software for T-cell receptor sequencing data analysis. *Nature Methods*, **10** :813–814.

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., & Chudakov, D. M., (2015). MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods*, **12** (5):380–381.

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., & Fire, A. Z., (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science Translational Medicine*, **1** (12):12ra23.

Boyd, S. D., Gaëta, B. A., Jackson, K. J., Fire, A. Z., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., & Collins, A. M., (2010). Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *The Journal of Immunology*, **184** (12):6986–92.

Brochet, X., Lefranc, M.-P., & Giudicelli, V., (2008). IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Research*, **36** (S2):W503–W508.

Brüggemann, M., Raff, T., & Kneba, M., (2012). Has MRD monitoring superseded other prognostic factors in adult ALL? *Blood*, **120** (23): 4470–4481.

Castro, R., Jouneau, L., Pham, H.-P., Bouchez, O., Giudicelli, V., Lefranc, M.-P., Quillet, E., Benmansour, A., Cazals, F., Six, A., Fillatreau, S., Sunyer, O., & Boudinot, P., (2013). Teleost fish mount complex clonal IgM and IgT responses in spleen upon systemic viral infection. *PLoS Pathogens*, **9** (1):e1003098.

Cavé, H., van der Werff Ten Bosch, J., Suciu, S., Guidal, C., Waterkeyn, C., Otten, J., Bakkus, M., Thielemans, K., Grandchamp, B., Vilmer, E., Nelken, B., Fournier, M., Boutard, P., Lebrun, E., Méchinaud, F., Garand, R., Robert, A., Dastugue, N., Plouvier, E., Racadot, E., Ferster, A., Gyselinck, J., Fenneteau, O., Duval, M., Solbu, G., & Manel, A.-M., (1998). Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia. *New England Journal of Medicine*, **339** (9):591–598.

Faham, M., Zheng, J., Moorhead, M., Carlton, V. E. H., Stow, P., Coustan-Smith, E., Pui, C.-H., & Campana, D., (2012). Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood*, **120** (26):5173–5180.

Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H., & Holt, R. A., (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research*, **19** (10):1817–24.

Gandemer, V., Pochon, C., Oger, E., Dalle, J.-H. H., Michel, G., Schmitt, C., Berranger, E., Galambrun, C., Cavé, H., Cayuela, J.-M., Grardel, N., Macintyre, E., Margueritte, G., Méchinaud, F., Rorhlich, P., Lutz, P., Demeocq, F., Schneider, P., Plantaz, D., Poirée, M., & Bordigoni, P., (2014). Clinical value of pre-transplant minimal residual disease in childhood lymphoblastic leukaemia: the results of the French minimal residual disease-guided protocol. *British Journal of Haematology*, **165** (3):392–401.

Gawad, C., Pepin, F., Carlton, V., Klinger, M., Logan, A., Miklos, D., Faham, M., Dahl, G., & Lacayo, N., (2012). Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood*, **120** (22):4407–17.

Giraud, M., Salson, M., Duez, M., Villenet, C., Quief, S., Caillault, A., Grardel, N., Roumier, C., Preudhomme, C., & Figeac, M., (2014). Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing. *BMC Genomics*, **15** (1):409.

Kerst, G., Kreyenberg, H., Roth, C., Well, C., Dietz, K., Coustan-Smith, E., Campana, D., Koscielniak, E., Niemeyer, C., Schlegel, P. G., Müller, I., Niethammer, D., & Bader, P., (2005). Concurrent detection of minimal residual disease (MRD) in childhood acute lymphoblastic leukaemia by flow cytometry and real-time PCR. *British Journal of Haematology*, **128** (6):774–782.

Kuchenbecker, L., Nienen, M., Hecht, J., Neumann, A. U., Babel, N., Reinert, K., & Robinson, P. N., (2015). IMSEQ – a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics*, **31** (18):btv309.

Laserson, U., Vigneault, F., Gadala-Maria, D., Yaari, G., Uduman, M., Vander Heiden, J. A., Kelton, W., Taek Jung, S., Liu, Y., Laserson, J., Chari, R., Lee, J.-H., Bachelet, I., Hickey, B., Lieberman-Aiden, E., Hanczaruk, B., Simen, B. B., Egholm, M., Koller, D., Georgiou, G., Kleinstein, S. H., & Church, G. M., (2014). High-resolution antibody dynamics of vaccine-induced immune responses. *Proceedings of the National Academy of Sciences*, **111** (13):4928–2933.

Lefranc, M.-P., (2011). IMGT, the International ImMunoGeneTics Information System. *Cold Spring Harbor Protocols*, **2011** (6): pdb.top115.

Logan, A. C., Zhang, B., Narasimhan, B., Carlton, V., Zheng, J., Moorhead, M., Krampf, M. R., Jones, C. D., Waqar, A. N., Faham, M., Zehnder, J. L., & Miklos, D. B., (2013). Minimal residual disease quantification using consensus primers and high-throughput IGH sequencing predicts post-transplant relapse in chronic lymphocytic leukemia. *Leukemia*, **27** :1659–1665.

Logan, A. C., Gao, H., Wang, C., Sahaf, B., Jones, C. D., Marshall, E. L., Buno, I., Armstrong, R., Fire, A. Z., Weinberg, K. I., Mindrinos, M., Zehnder, J. L., Boyd, S. D., Xiao, W., Davis, R. W., & Miklos, D. B., (2011). High-throughput VDJ sequencing for quantification of minimal residual disease in chronic lymphocytic leukemia and immune reconstitution assessment. *Proceedings of the National Academy of Sciences of the United States of America*, **108** (52):21194–21199.

Logan, A. C., Vashi, N., Faham, M., Carlton, V., Kong, K., Buno, I., Zheng, J., Moorhead, M., Klinger, M., Zhang, B., Waqar, A., Zehnder, J. L., & Miklos, D. B., (2014). Immunoglobulin and T cell receptor gene high-throughput sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-transplantation relapse and survival. *Biology of Blood and Marrow Transplantation*, **20** (9):1307–1313.

Market, E. & Papavasiliou, F. N., (2003). V(D)J recombination and the evolution of the adaptive immune system. *PLoS Biology*, **1** (1):E16.

Niklas, N., Pröll, J., Weinberger, J., Zopf, A., Wiesinger, K., Krismer, K., Bettelheim, P., & Gabriel, C., (2014). Qualifying high-throughput immune repertoire sequencing. *Cellular Immunology*, **288** :31–38.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B., (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, **14** (5):R51.

Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., & Chain, B., (2013). Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, **29** (5):542–550.

Tonegawa, S., (1983). Somatic generation of antibody diversity. *Nature*, **302** (5909):575–581.

van Dongen, J. J. M., Szczepański, T., & Adriaansen, H. J. Immunobiology of leukemia. In ES, H., TA, L., & M, G., editors, *Leukemia*, pages 85–130. WB Saunders, Philadelphia, 7th edition, (2002).

van Dongen, J. J. M., Langerak, A. W., Brüggemann, M., Evans, P. A. S., Hummel, M., Lavender, F. L., Delabesse, E., Davi, F., Schuuring, E., García-Sanz, R., van Krieken, J. H. J. M., Droese, J., González, D., Bastard, C., White, H. E., Spaargaren, M., González, M., Parreira, A., Smith, J. L., Morgan, G. J., Kneba, M., & Macintyre, E. A., (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 concerted action BMH4-CT98-3936. *Leukemia*, **17** (12):2257–317.

van Dongen, J. J., van der Velden, V. H., Brüggemann, M., & Orfao, A., (2015). Minimal residual disease (MRD) diagnostics in acute lymphoblastic leukemia (ALL): need for sensitive, fast and standardized technologies. *Blood*, **125** (26):3996–4009.

Warren, R. L., Nelson, B. H., & Holt, R. A., (2009). Profiling model T-cell metagenomes with short reads. *Bioinformatics*, **25** (4):458–64.

Warren, R. L., Freeman, J. D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J. R., & Holt, R. A., (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*, **21** (5):790 –797.

Weinstein, J. A., Jiang, N., White, R. A.3rd, Fisher, D. S., & Quake, S. R., (2009). High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324** (5928):807–10.

Weng, W.-K., Armstrong, R., Arai, S., Desmarais, C., Hoppe, R., & Kim, Y. H., (2013). Minimal residual disease monitoring with high-throughput sequencing of T cell receptors in cutaneous T cell lymphoma. *Science Translational Medicine*, **5** (214).

Wu, D., Sherwood, A., Fromm, J. R., Winter, S. S., Dunsmore, K. P., Loh, M. L., Greisman, H. A., Sabath, D. E., Wood, B. L., & Robins, H., (2012). High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. *Science Translational Medicine*, **4** (134):134ra63–134ra63.

Yang, X., Liu, D., Lv, N., Zhao, F., Liu, F., Zou, J., Chen, Y., Xiao, X., Wu, J., Liu, P., Gao, J., Hu, Y., Shi, Y., Liu, J., Zhang, R., Chen, C., Ma, J., Gao, G. F., & Zhu, B., (2014). TCRklass: A new k-string-based algorithm for human and mouse TCR repertoire characterization. *Journal of Immunology*, **194** (1).

Ye, J., Ma, N., Madden, T. L., & Ostell, J. M., (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, **41** :W34–W40.

Yousfi Monod, M., Giudicelli, V., Chaume, D., & Lefranc, M.-P., (2004). IMGT/JunctionAnalysis: the first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics*, **20** (S1):i379–85.