Author's response to reviews

Title: Whole exome sequencing of microdissected splenic marginal zone lymphoma: a study to discover novel tumor-specific mutations.

Authors:

Jan Peveling-Oberhag (jan.peveling-oberhag@kgu.de)
Franziska Wolters (franziwolters@web.de)
Claudia Döring (C.Doering@em.uni-frankfurt.de)
Dirk Walter (Dirk.Walter@kgu.de)
Ludger Sellmann (LSellmann@gmx.net)
René Scholtysik (rene.scholtysik@uni-due.de)
Marco Lucioni (m.lucioni@smatteo.pv.it)
Max Schubach (maxschubach@web.de)
Marco Paulli (m.paulli@smatteo.pv.it)
Saskia Biskup (saskia.biskup@cegat.de)
Stefan Zeuzem (Zeuzem@em.uni-frankfurt.de)
Ralf Küppers (ralf.kueppers@uk-essen.de)
Martin-Leo Hansmann (martin.l.hansmann@em.uni-frankfurt.de)

Version: 3
Date: 15 July 2015

Author's response to reviews: see over
Dear Sir,

Our manuscript has been reviewed by you and three experts in the field. The comments of the reviewers were most helpful to improve the paper and we have carefully revised the manuscript according to all their suggestions.

All changes in the revised manuscript are underlined, and a detailed point-by-point response to the specific comments of the reviewers is attached. Among other modifications we now provide additional sequencing analyses to confirm the somatic origin of the SMYD1 mutations. Furthermore, we now elaborate possible functional consequences of the SMYD1 mutations, including the impact on the protein structure using a murine model of the crystalline protein structure. We feel that the manuscript has been significantly improved by the introduced modifications.

The data of the manuscript have not been published previously and are not under consideration for publication elsewhere. We would be grateful for consideration of the present manuscript to be published in *BMC Cancer*.

Jan Peveling-Oberhag, M.D.
Corresponding author
Point-by-point response

Reviewer 1:

1. The introduction is missing key references concerning recent papers in this area, focusing on KLF2 from and Italian and UK group, but also a large clinical study published in Clinical Cancer Research (which also performs bioinformatics analysis of deep-sequencing data to provide some preliminary insights into sub-clonal diversity).

We now integrated these important recent data in the introduction and discussion section of the revised manuscript [1-3] (page 3, line 20ff; page 12, line 17ff).

Furthermore, we reanalysed our WES data specifically for KLF2 mutations. The bioinformatic SNV-detection algorithm did not detect any mutations in this area. Also manual review of the region of interest did not detect KLF2 mutations. However, large areas of this gene showed poor or no coverage, although the enrichment design included specific baits for this region. A possible explanation for this drop-out in sequencing coverage might be the high GC content of KLF2. Also, the Exome Aggregation Consortium database of the Broad Institute shows that within >60,000 sequenced exomes there is largely no or minimal coverage of KLF2. Therefore, we might have failed to detect actual KLF2 mutations in our cases (false negatives). We state these findings and acknowledged this shortcoming of our study in the results section and the discussion of the revised manuscript (page 8, line 17ff; page 12, line 17ff).

2. The discussion states that their approach to micro-dissecting samples provided better insight into clonal evolution, but there technical design of the experiment does not support this, i.e. such low sequencing depth.

We agree with the reviewer that our technical design is not ideally suited to detect sub-clonal changes and much less clonal evolution. A targeted re-sequencing approach with high sequencing depth as used by Parry et al. provides much more accurate data on minor allele frequencies. Low global coverage is a general downside of WGS or WES approaches and detection of sub-clonal events may only partly be enabled by
high tumor cell purity. Nevertheless, we observed SNVs with allele frequencies close to 0.5 and several other positions with significantly lower allele counts, not only in the WES approach but also in the confirmatory analysis using pyrosequencing. These specific regions of interest were covered >40-fold. Therefore, we speculated on a possible sub-clonal origin. We now rephrased and shortened the section about sub-clonality in the discussion section. We now clearly point out that the WES method used is not well suited to reliably detect such aberrations (page 11, line 7ff and 17ff).

3. We know certain genes are false positives in this type of analysis using MutSig approaches, but the authors do not mention this.
We agree that there is likely a high rate of false positives as well as false negatives (e.g. GC-rich regions such as KLF2) to be expected in the WES approach. This has now been acknowledged in the discussion section of the revised manuscript (page 12, line 25ff and page 13, 1ff).

4. I cannot see Figure 1, so cannot assess it. Is the title of this figure correct, as I am not sure it shows ‘coding complexity’?
We rephrased the title to: “Distribution of single nucleotide variants in the coding SMZL genome” (page 21, line 3ff).
The figure has been uploaded in a different format to allow the reviewer a better assessment of the data.

5. Is the MYD88 mutation in the dup(3q) case duplicated? You would expect this, and it would be interesting to know.
This is indeed interesting. When reassessing this region in the dup(3q) case we did not find the MYD88 794T>C mutation duplicated. The mutation was called in only one third of the reads (14 of 50), so likely the mutation is located on the non-duplicated region of chromosome 3. We added this observation to the results section of the revised manuscript (page 8, line 5ff).
6. **What are the coverage stats for certain genes, as we know NOTCH2 is poorly captured in WES studies?**

   We added a table of gene specific coverage stats to the revised version of the manuscript (Table S5, page 27). Indeed, as mentioned above KLF2 shows a rather low coverage with a mean of 13.3 (range 1-45). However, NOTCH2 shows relatively high coverage with a mean of 66.8 (range 35-146) in our study and also general coverage statistics for NOTCH2 in global WES studies are rather good (Exome Aggregation Consortium database of the Broad Institute shows a mean coverage of 67.1). The data have now been integrated in the revised results section of the manuscript (page 8, line 22ff).

7. **The authors should consider providing KLF2 data on their cohort, as this is the most important gene in this disease.**

   As stated above, KLF2 is covered only insufficiently in our WES. We decided against re-sequencing for KLF2 mutation in our cohort, firstly, as we did not detect any mutations in our discovery cases (possibly due to above mentioned technical reasons). Secondly, we felt that any outcome in our small cohort would not alter the view on the significance of this mutation when compared to the large British and Italian case numbers.

8. **I think the discussion is ‘over sold’ and needs to be toned back to be a better reflection of actually what they add to the field. The sub-clonal section particularly is not discussed in the context of the quality and quantity of the sequencing data they have.**

   We addressed this issue. Please see #2.

9. **The authors mention a minimum coverage of 20 and a novel VAF of 10%, so 2 reads for calling a variant is very low and lightly to include high levels of false positives. Furthermore, only 66% of the exome was captured at this depth, so the authors are missing a lot here.**

   We agree with the reviewer that the minimum coverage requirement may lead to a significant loss of detected mutations (false negatives). Concerning the false positive rate, all positions meeting the above mentioned criteria were reviewed manually by two different observers to minimize false positives. This has been clarified in the
methods section and acknowledged in the discussion of the revised manuscript (page 7, line 6ff; page 10, line 17ff).

10. **Are these (SMYD) validation variants somatically-acquired?** Is this frequency of mutations highly then you would expect based on background rates? Have these mutations been reported in the extensive exome sequencing data published already?

We now performed additional Sanger sequencing experiments with the germline DNA of patients with the SMYD1 mutation. In the discovery cases germline material was already available (CD3 MACS-sorting) and used also for WES. To collect germline material of the validation cases we used archived GI biopsy material without lymphoma manifestation (gastric biopsy) where available. For one case we microdissected the splenic capsule of the resected spleens to gather non-neoplastic cells. We did not detect any germline SMYD1 mutations in the genomic regions of interest, neither in the discovery case nor in the validation cohort. These findings were added to the results section of the revised manuscript (page 9, line 15ff).

Furthermore, we compared our findings to three NGS databases: 1000 genome project (Sanger Institute / Broad Institute), Exome Aggregation Consortium (60,706 exomes; Broad Institute), Catalogue of somatic mutations in cancer (COSMIC, Sanger Institute). The database content is partly overlapping. Nevertheless, the SMYD1 mutations found in our current SMZL cohort are not described in one of the mentioned platforms. SMYD1 seems rather well conserved with only 189 missense or loss-of-function variants throughout the gene. COSMIC database contains 170 unique cancer samples with SMYD1 mutations out of 24,615 total samples (none of these at one of the three positions we found mutated in the current analysis). Mutations have been discovered in various cancer types (e.g. gastric-, hepatocellular-, bladder-, renal cell carcinoma, melanoma or glioma). These additional analyses have been included in the revised version of the manuscript (page 9, line 16ff).
Reviewer 2:

1. **The functional relevance of SMYD1 mutation should be explored**

   To further look into the functional relevance of detected SMYD1 variants we used different tools:

   1. Polyphen and SIFT algorithm predict possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations. The two algorithms predicted the found mutations to be either possibly damaging and damaging, respectively.

   2. Literature research showed that the SET domain is the crucial active part of the SMYD1 protein acting through histone modification [4]. The homologous domain in SMYD2 likely represses the activity of p53 and RB through methylation [5, 6]. Both currently detected exonic mutations are located in the SET domain.

   3. An analysis using Universal Protein Resource (UniProt) by the European Bioinformatics Institute [7] shows that one of the found mutations (c.836G>T; p.C279F) is located at one of the four zinc binding sites of the catalytic centre of SMYD1.

   4. We used the crystal structure of the murine SMYD1 protein (retrieved from the Protein Data Bank; PDB-ID: 3N71) and localized the 2 coding mutation. Interestingly, the p.S321Y mutation lies within the functional pocket of the protein.

   5. The position in the 3’ UTR of SMYD1, where we found one of the mutations, is a predicted target of miR28, a recently described potent tumor suppressor in B-cell lymphomagenesis.

These additional analyses and one additional supplementary figure (crystalline structure of the protein) have been included in the revised version of the manuscript (page 10, line 2ff; page 13, line 15ff).

2. **Clinical correlation of these findings may give more relevance to these findings**

   We do have comprehensive clinical data available on the included cases. However, the low case number combined with the low frequency of detected variants (SMYD1: 3 out of 10 cases, MYD88 3 out of 24 cases) does not allow for a reasonable statistic correlation with clinical parameters. Therefore, we did not include clinical correlation in the current manuscript. However, we are planning to analyse SMYD1 mutations in
a larger cohort in the future to be able to correlate with clinical features of these patients.

3. **Proportion of neoplastic cells should be somehow scored**

   We used saved images of the computer supported microdissection of the two discovery cases to estimate tumor cell content. Neoplastic cells within the microdissected areas were counted in those representative images and proportions (87% and 89%) are now mentioned in the revised version of the manuscript (page 5, line 6ff). However, only a limited number of images were taken during microdissection for each case. Therefore, the neoplastic cell proportion remains an approximation.

4. **The study does not confirm KLF2 mutations, now described to be the most frequent mutational event in SMZL. A specific search for these mutations would consolidate the data here provided.**

   We reanalysed our WES data specifically for KLF2 mutations. The bioinformatic SNP-detection algorithm did not detect any mutations in this area. Also manual review of the region of interest did not detect KLF2 mutations. However, large areas of this gene showed poor or no coverage, although the enrichment design included specific baits for this region. KLF2 is known to have poor coverage in whole exome sequencing approaches. A possible explanation for this drop-out in sequencing coverage might be its high GC content. Therefore, we might have failed to detect actual KLF2 mutations in our cases. We acknowledged this shortcoming in our discussion section of the revised manuscript (page 8, line 17ff; page 12, line 17ff). See also #1 of Reviewer 1.
Reviewer 3:

1. The novelty of the paper is limited. Exome sequencing of unselected spleen tissues gave essentially the same results when applied to larger cohorts of patients. Novelty would be higher if the authors performed multiregional sequencing of SMZL cells microdissected from different regions of the spleen. Such approach might reveal the clonal architecture and evolution history of SMZL.

   We have to acknowledge that general mutation detection (apart from mutations in the SMYD1 gene) did not differ significantly from studies with non-microdissected tissues. However, we gathered some preliminary evidence on possible subclonal alterations which could have been missed by other studies.

   Sequencing of different microdissected histological regions of SMZL would be a very interesting approach to gain further insight into lymphoma subclonality. As also mentioned by reviewer 1, ultra-deep re-sequencing of circumscribed genetic regions is superior to the WES approach applied in the current study in quantifying subclonal evolution. Panel-sequencing of microdissected intra-tumoral regions could be a promising follow-up project of the current analysis.

2. Page 3, line 8. The statement "The molecular pathogenesis of SMZL remains unclear" should be rephrased. Indeed, a number of recently published studies have disclosed recurrently mutated genes in SMZL, including NOTCH2 and KLF2.

   We rephrased the sentence in the revised version of the manuscript acknowledging recent data on NOTCH2 and KLF2 (page 3, line 8ff). Please see also #1 of Reviewer 1.

3. Page 6, line 24. "Somatic variants were called if the allele frequency of the normal tissue was smaller 0.2 and the delta between tumor and normal frequency was at least 0.1." The method the authors used to call somatic variant does not seem as stringent as it would be recommended. Indeed, the variant allele frequency cut-off in the normal tissue is high (20%).

   We agree with the reviewer that the algorithm to detect somatic variants has shortcomings in stringency. However, all positions meeting the above mentioned criteria were reviewed manually by two different observers to exclude false positives.

   We used this approach to reduce the number of false negative findings. This has been
clarified in the methods section and acknowledged in the discussion of the revised manuscript (page 7, line 6ff; page 10, line 17ff).

4. **The authors assessed the recurrence of specific mutations in the validation set of SMZL cases. This approach has advantages if the mutation is known to recurrently affect a hotspot (i.e. the MYD88 L265P variant). However, in the setting of genes without known mutation hotspots, variants might be dispersed across the entire coding region. Therefore the entire coding region plus splice sites of the discovered genes should be analyzed in the validation panel.**

We agree with the assessment of the reviewer. To account for the mentioned problem of the validation approach we used Sanger sequencing of either whole exonic regions (all 10 exons of SMYD1) or important regions of interest (PEST domain of NOTCH2) in cases with available fresh tissue. Additionally we utilized pyro-sequencing of known or possible hot spots (e.g. MYD88 L265P). The advantage of the hot spot sequencing was the superior applicability to degraded FFPE tissue, as fresh tissue was not available in all cases. This technical limitation has been acknowledged in the manuscript on page 9, line 4ff of the revised manuscript.

5. **The somatic origin of SMYD1 variants should be confirmed by testing the paired normal DNA.**

We agree with the reviewer and have now performed additional Sanger sequencing experiments with the germline DNA of patients with the SMYD1 mutation.

In the discovery cases germline material was already available (CD3 MACS-sorting) and used also for WES. To collect germline material of the validation cases we used archived GI biopsy material without lymphoma manifestation (gastric biopsy) where available. For one case we microdissected the splenic capsule of the resected spleens to gather non-neoplastic cells. We did not detect any germline SMYD1 mutations in the genomic regions of interest, neither in the discovery case nor in the validation cohort. These findings were added to the results section of the revised manuscript (page 9, line 15ff).

Furthermore, we compared our findings to three NGS databases: 1000 genome project (Sanger Institute / Broad Institute), Exome Aggregation Consortium (60,706 exomes; Broad Institute), Catalogue of somatic mutations in cancer (COSMIC, Sanger Institute). The database content is partly overlapping. Nevertheless, the SMYD1
mutations found in our current SMZL cohort are not described in one of the mentioned platforms. SMYD1 seems rather well conserved with only 189 missense or loss-of-function variants throughout the gene. COSMIC database contains 170 unique cancer samples with SMYD1 mutations out of 24,615 total samples (none of these at one of the three positions we found mutated in the current analysis). Mutations have been discovered in various cancer types (e.g. gastric-, hepatocellular-, bladder-, renal cell carcinoma, melanoma or glioma). These additional analyses have been included in the revised version of the manuscript (page 9, line 16ff).

6. Given the relatively high prevalence of MYD88 mutations and the relative scarcity of NOTCH2 mutations, the authors should rule out the inclusion in the study cohort of disorders looking alike SMZL (i.e. LPL).

The rate of MYD88 mutations was not significantly higher than in most of the other WES approaches, but we agree with the reviewer that morphologically similar lymphoma subtypes such as LPL should be carefully avoided. All included cases were typical SMZL cases and reviewed independently by at least 2 pathologists (most were reviewed in a panel of pathologists). Nevertheless, following the comment of the reviewer we re-reviewed those cases with MYD88 mutations. However, the review did not lead to a change of diagnosis in these cases. The re-review has been mentioned in the revised version of the manuscript (page 4, line 22ff).
References:


