Article Title: *IGHV* sequencing reveals acquired N-glycosylation sites as a clonal and stable event during follicular lymphoma evolution

Short Title: **Fate of N-glycosylation of *IGHV* in FL progression**

Mariette Odabashian, [1] Emanuela Carlotti, [1] Shamzah Araf, [1] Jessica Okosun, [1] Filomena Spada, [1] John Gribben, [1] Francesco Forconi, [2] Freda K Stevenson, [2] Mariarita Calaminici, [1, *] and Sergey Krysov [1, *]

[1] Centre for Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, London, UK. [2] Cancer Sciences Division, Somers Cancer Sciences Building, University of Southampton, Southampton, UK.

[*] M.C. and S.K. contributed equally to this study.

Correspondence: S. Krysov, Centre for Haemato-Oncology, Barts Cancer Institute, John Vane Science Centre, Queen Mary University of London, London, EC1M 6BQ, UK; email: s.krysov@qmul.ac.uk; phone number: +44 (0)207 882 3823; fax number: +44 (0)207 882 3881.

Word count: Abstract: 250, Text: 3995

7 Figures, 4 Tables, 24 References

Key Points

- N-glycosylation sites are acquired early in disease and persist during tumour progression despite therapy.
- Scarcity of N-glycosylation sites-negative subclones and their loss during progression suggest positive clones expand preferentially.

Abstract

Follicular lymphoma B cells undergo continuous somatic hypermutation (SHM) of their immunoglobulin variable region genes, generating a heterogeneous tumour population. SHM introduces DNA sequences encoding N-glycosylation sites Asparagine-X-Serine/Threonine (N-gly sites) within the *V*-region that are rarely found in normal B cell counterparts. Unique attached oligomannoses activate B cell receptor signalling pathways following engagement with calcium-dependent lectins expressed by tissue macrophages. This novel interaction appears critical for tumour growth and survival. To elucidate the significance of N-gly site presence and loss during ongoing SHM, we tracked site behaviour during tumour evolution and progression in a diverse group of patients through next-generation sequencing. A hierarchy of subclones was visualised through lineage trees based on SHM semblance between subclones and their discordance from the germline sequence. We observed conservation of N-gly sites in >96% of subclone populations within and across diagnostic, progression and transformation events. Rare N-gly-negative subclones were lost or negligible from successive events in contrast to N-gly-positive subclones which could additionally migrate between anatomical sites. Ongoing SHM of the N-gly sites resulted in subclones with different amino acid compositions across disease events, yet the vast majority of resulting DNA sequences still encoded for an N-gly site. The selection and expansion of only N-gly-positive subclones is evidence of the tumour cells dependence on sites despite the changing genomic complexity as the disease progresses. N-gly sites were gained in the earliest identified lymphoma cells, indicating they are an early and stable event of pathogenesis. Targeting the inferred mannose-lectin interaction holds therapeutic promise.

Introduction

Follicular lymphoma (FL) is a biologically and clinically heterogeneous disease that remains incurable. Although the majority of patients follow an indolent course, a high-risk group are prone to early progression or transformation to aggressive lymphoma associated with a dismal prognosis. For these patients, current therapies are suboptimal and uncovering changes occurring early during disease development is essential to improving prognosis.

Despite the loss of one immunoglobulin allele through the t14;18 translocation [1] and ongoing somatic hypermutation (SHM) of the immunoglobulin heavy-chain variable region gene (*IGHV*) that can introduce crippling *IGV* mutations, all detectable tumour subclones retain functional expression of the surface immunoglobulin throughout disease, resulting in thousands of tumour subclones displaying distinct but clonally related *IGHV* sequences. This retention suggests a tumour dependence on signalling through the B cell receptor (BCR). Through SHM, replacement mutations introduce amino acid sequence motifs consisting of Asparagine (N)-X-Serine/Threonine (S/T), where X can be any amino acid except proline. [2-4] These sequences are known as N-glycosylation (N-gly) sites and are found in >90% of FL cases. [5, 6] N-gly sites are rarely found in normal B cells, [7] indicative of a pathogenic function. Unusual glycans terminating with high-mannose attach to sites and activate BCR signalling pathways following engagement with lectins. [8-12] This novel interaction represents a critical mechanism by which tumour cells survive in the germinal centre (GC), accumulating mutations of epigenetic modifiers early during FL pathogenesis. [13, 14]

The behaviour of N-gly sites during disease evolution and progression has been investigated by *IGHV* cloning technique in a number of FL cases [5] and in one case of contiguous FL and in situ follicular neoplasia (ISFN). [15] These studies have indicated conservation of acquired N-gly sites within identified clones. However, clone numbers were limited in these studies, underrepresenting the extent of intraclonal diversity. Furthermore, as analysis has been restricted to a single disease event, behaviour of N-gly sites over time has not been addressed and would be critical in determining their role in disease initiation and progression. To address this requires comprehensive *IGV* analysis of the clonal repertoire taken from subsequent (temporal) biopsies, ranging from a relatively early time point in disease manifestation (e.g. diagnosis) to a time point at which the disease has become genetically and clinically distinct (e.g. relapse and transformation). As SHM continues during disease progression and transformation, the stepwise process can be visualised through lineage trees rooted to a putative non-malignant germline *IGV* sequence, making them an important tool in B cell evolutionary studies.

Our goal was to investigate the behaviour of N-gly sites during the disease course. We analysed the incidence and maintenance of sites within the tumour clones of six patients taken at different time points of disease. This included analysis of events from different anatomical sites. This is the first study that has analysed the relationship between FL progression and N-gly sites in patients who have undergone different lines of therapy and presented with different clinical courses, reflecting the heterogeneous nature of the disease.

We found that N-gly sites are acquired within early FL clones and are retained in the intraclonal population despite ongoing SHM. A striking observation is that sites are a universal determinant of both cell expansion and cell fate, as evidenced by the low frequency of N-gly site-negative subclones in and across diagnostic, progression and transformation events and their disappearance in subsequent events.

Methods

Methods

Three individuals with FL were selected on the availability of genomic DNA derived from sequential tumour lymph node biopsies that had previously undergone somatic variant profiling to reveal a 'sparse' or 'rich' disease evolution pattern based on degree of genetic semblance. [13] Patient's 1 and 3 were categorised as 'sparse' and patient 2 categorised as 'rich'. Samples were selected based on detection of a clonal *IGHV* (*IGHVDJ*) rearrangement through Sanger sequencing (Supplemental Methods). In total, 8 samples were selected, all carrying an IgH-VH3 rearranged major tumour clone (major clone). All samples were obtained after written informed consent in accordance with the Declaration of Helsinki and the London Research Ethics Committee. ~50ng of *IGHV* genomic DNA amplicons prepared using JH consensus and VH3-FR1 primers were sent for 2x250bp paired end sequencing using the Miseq Illumina platform (Genewiz, NJ). As primers bind within the FR1 and JH regions, a portion of these regions were absent from the sequencing data. The sequential steps involved in the analysis of Illumina reads and identification of tumour related subclones are detailed in the Supplemental Methods. Additional tumour related reads covering the *IGHV* gene for 2 patients over different disease events were available from our collaborator (patients 4 and 5) and were produced using Roche 454 Life Sciences Genome Sequencer FLX. [16] Additional raw *IGHV* data files produced from the MiSeq platform were obtained from the NCBI database (BioProject PRJNA240336) for patient 6. [17] Clones were analysed for acquired N-gly sites using the NetN-glyc 1.0 server. Lineage trees based on the SHM profiles of clones were generated using IgTree, [18] detailed in the Supplemental Methods.

Statistical analyses

Two-way ANOVA was performed using GraphPad Prism (GraphPad Software, La Jolla, CA).

Results

High throughput sequencing analysis of tumour related subclones

Sequencing metrics for patients 1-3 are found in Supplemental Table 1. We generated 0.81 to 1.17 million paired-end reads/sample (average 1.09 million) (Supplemental Table 2). In total, we identified 0.12 to 0.46 million (average 0.29 million) *VDJ* junctions per sample. The major clone was identified as being the dominant *VDJ* rearrangement in the sample (Table 1) and tumour related reads were identified as described in the Supplemental Methods. The number of unique subclones that reads encoded for is detailed in Table 1. To ensure detection of all tumour subclones in the different samples we utilized VH3 family oligonucleotides in the single sequencing run rather than tumour specific primers. Therefore, contaminating sequences from normal B cells were observed for each sample and the number of unique *VDJ* rearrangements/sample are stated in in Table 1. Patients 4-6 sequencing data in Table 1 was extracted from the original articles. [16, 17] The relatively greater number of unique *VDJ* rearrangements detected for patient 6 is due to the sequencing approach amplifying all VH families. [17] There is a heterogeneous level of contaminating B cells, as indicated by the percentage of merged reads expressing the dominant tumour rearrangement, ranging from 53.3% to 99.03% (Table 1 and Supplemental Table 2).

Site conservation is a universal feature of the tumour clonal population

We sequenced the *IGHV* gene in samples obtained at sequential time-points of FL in six patients and interrogated the derived tumour sequence for the acquisition of N-gly sites. Details regarding patient samples can be found in Supplemental Table 3. All major clones identified across samples contained

one or more N-gly sites (Table 2). With the exception of patient 5, N-gly sites were conserved across disease events, despite patients undergoing several lines of therapy in between biopsies (Supplemental Table 3). For patient 1 and the 4th N-gly site of patient 2, sites were conserved in transformation events through non-silent mutations that impacted the amino acid sequence (e.g. NFS>NVS). For the remaining sites in patient 2 and sites in patient 3 and 4, the amino acid sequences were conserved across disease events. Conservation of sites is also supported in our extension cohort of serial FL and transformed samples from patients A-E that underwent *IGHV S*anger sequencing (Supplemental Table 4). Patient 5 and 6 sequential samples were derived from different anatomical sites (Supplemental Table 3). For patient 5, the two disease events have distinct N-gly sites; NFS in the CDR1 region (1st relapse event) and NLT in the FR3 region (3rd relapse event). For patient 6, all events contain the same N-gly site and amino acid motif (NGS) (Table 2). To elucidate whether N-gly site acquisition is a clonal event, we interrogated the subclone population by next-generation sequencing. For patients containing one N-gly site in their major clone, ≥97% of the subclone population within and across disease events maintained the site (Table 2). For patients 2 and 4, which had multiple N-gly sites, no subclone with the complete absence of N-gly sites was detected. For patient 2, the 1st, 2nd and 3rd sites were conserved in >96% of subclones across events. The 4th site was conserved in 97.2% of clones in the 1st relapse sample and in 82% and 85% of subclones in the 3rd relapse and transformation samples, respectively. Interestingly, for all patients no further N-gly site accumulated within or across events that were not found in the major clone. This infers that site acquisition is a conserved event.

SHM diversity within the N-glycosylation site indicates a selective retention of site-positive subclones

Table 3 highlights the number of unique subclones that have a different sequence in the N-gly site region compared to the major clone of the disease event. As the N-gly site is encoded by nine nucleotides, these subclones differ from the major clone by at least one nucleotide within the N-gly site region. There is wide variation in percentage of affected subclones between patients, ranging from 0 to 58.41% of the total subclone population, indicating the (largely) random targeting of SHM within the variable region. For patients 1, 2 (sites 1-3), 3, 4 (site 1) and 5, the majority of affected subclones across disease events maintain the N-gly site, indicating a positive selection (Figure 1a). Analysis of the codon sequences of these positive subclones across patients and events revealed that N-gly sites are retained through either synonymous mutations or non-synonymous mutations. The profiling of these subclones from patient 3 is used to highlight these two means of N-gly site retention in Figure 1b.

For patient 2, the 4th N-gly site was absent in the majority of affected subclones in the 3rd relapse and transformation events, yet remaining N-gly sites may be supporting their survival and expansion as indicated by their high percentage in the subclone population. While in patient 6 the affected subclones in the first two events were mostly N-gly site positive, the affected subclones of the relapsed tFL event were predominantly site negative. As this is a relatively late disease event compared to the other patients, the N-gly site may have become redundant in promoting tumour survival. However, it is important to point out that these negative subclones make up 2.7% of the total subclone population (Table 3). Furthermore, site-negative subclones only making up ≤1% of the total count number in samples expressing only one N-gly site (Supplemental Table 5), indicating they are a minor component of the tumour bulk.

N-glycosylation sites in distinct anatomical sites

The distinct anatomical sites for patients 5 and 6 serial samples make them important in studies regarding the genealogy of N-gly sites. For both patients, serial disease events were derived from the

same precursor B cell, as evidenced by a shared *VDJ* rearrangement and t(14;18) translocation. For patient 5, the two events have distinct N-gly sites (Table 2) whereas for patient 6, all events contain the same N-gly site and amino acid motif (NGS). When comparing the subclones of patient 5, we observe a clear discordance in the SHM pattern of the two temporal populations and how this translates into a highly distinct amino acid sequence (Supplemental Figure 1). This suggests that for patient 5, there was early divergence of the precursor tumour cell before N-gly site acquirement whereas for patient 6, the precursor cell diverged after acquiring the N-gly site (Figure 2). Patient 5 demonstrates that N-gly acquirement may not be an event of an early divergence evolution model, instead occurring in anatomical site-specific ancestral cells that have undergone unique SHM processes. However, acquirement occurs early in these site-specific cells, as illustrated by the presence of N-gly sites in 97.14% and 99.19% of unique subclones in the 1st and 3rd relapse events, respectively (Table 2).

N-glycosylation site positive subclones are important in disease progression and migration between anatomical sites.

As described above and with the exception of patient 5, N-gly sites are conserved in the clonal population across disease events. When we compared subclone populations, the majority of subclones for each disease event are unique, highlighting the inter-tumour heterogeneity generated through SHM (Figure 3). The number of shared subclones make up 0.03-27.5% of the total tumour subclones identified across disease events. These subclones survive for years, as indicated by the intervals between temporal biopsy acquirement (Supplemental Table 3).

Analysis of shared subclones revealed they are all N-gly site-positive. Patient 3 was an exception as 1.5% of the shared clones were negative (n=20). Interestingly, most of these negative clones make up a higher percentage of the total tumour count in the successive disease event, suggesting that they confer an advantage (Supplemental Table 6). The lack of shared N-gly site-negative subclones in all other patients indicates that progression subclones are dependent on N-gly sites for their long-term survival. Analysis of shared subclones in patients 1 and 2, in which the amino acid composition of N-gly sites changes between disease events in both the major clones and subclone populations (Table 2), reveals that subclones giving rise to transformation tumours were already pre-existent as minor subclones in earlier events, gaining clonal dominance following therapy to generate the transformation tumours. This subclone plasticity relies on the conservation of N-gly sites, indicating the important role sites provide subclones involved in disease progression.

Subclones were also shared between biopsy sites, making up 0.4% and 0.3% of the overall subclone population across all disease events for patients 5 and 6, respectively. Similar to the other patients, shared subclones of both patients were all N-gly site positive, with patient 5 subclones containing the site of the first event (NFS motif in the CDR2 region). This indicates that migratory subclones require site presence and could represent a tumour cell feature which is critical for establishing disease in new locations, however this requires investigation in a larger cohort. The lack of shared subclones containing the N-gly site of the second disease event in patient 5 indicates that this disease event did not arise due to a pre-existing minor subclone in the first event gaining clonal dominance and repopulating the tumour at another site. This is in contrast to the subclone plasticity we observe in patients 1 and 2, described above.

N-glycosylation sites are acquired early in disease evolution

We can gain insight into tumour evolution by analysing the degree of SHM in each subclone. The range of SHM for each patient is indicated in Table 4, in which the least and most mutated subclones for

each disease event (compared to their germline sequence) were identified by the IMGT High V-QUEST program. The % difference in homology between the least and most mutated subclones ranged from 2.0 to 21.7%.

For patients 1, 4 and 6, N-gly sites were acquired within their least mutated subclones in all their disease events. For patient 6, acquirement of the CDR3 located site was observed after only four nucleotide substitutions (97.8% sequence homology to germline *V* gene) (Table 4). However, for the least mutated subclone of the transformed event for patient 3, the N-gly site isn't acquired despite the subclone harbouring a relatively greater number of point mutations. Despite this heterogeneity, N-gly sites are conserved once acquired despite ongoing SHM, as evidenced in the most mutated subclones of all patients. Therefore, subclone selection is based on conserving N-gly sites in spite of active mechanisms which have the potential to disrupt this.

Lineage trees specifically based on *VDJ* sequences were used to visualise the evolutionary intraclonal hierarchy.[18] For patients 4 and 5, the complete hierarchy can be visualised in lineage trees (Figures 4 and 5 and Supplemental Figure 2).

For patient 4, the earliest experimentally derived subclones (identified as filled circles closest to the germline Ig sequence at the top of the tree) are N-gly site-positive. While patient 4 does not have any truly negative subclones, several subclones lose at least one N-gly site. Some of these subclones are observed to undergo further SHM and re-acquire the lost site (Figure 4), giving rise to several further clones. This is in contrast to patient 5 in which the loss of the single N-gly site results in the subclone not undergoing further diversification or expansion (Figure 5). One N-gly site-negative subclone in the 1st relapse event is placed high in the tree and only differs from the germline sequence by 5 bases. This subclone corresponds to the least mutated subclone (98% homology) highlighted in Table 4, indicating this clone never acquired the N-gly site. The other site-negative subclones were descendants of site-positive clones because of ongoing SHM. As these clones are lost from progression samples, we can infer their elimination.

N-glycosylation site-negative clones arise from further SHM of site-positive clones

With greater numbers of N-gly site-negative subclones owing to an increase in overall subclones, patients 1, 3 and 6 lineage trees give a more comprehensive insight into the behaviour of site-negative clones in the tumour hierarchy (Figure 6). As patient 2 did not have any truly negative subclones, the analysis was omitted. However, these negative subclones represent a minority within the heterogeneous population. For patient 1, negative clones represented 1.7% and 1.8% of the subclone population in diagnosis and transformation events, respectively. For patient 3, negative clones found in 2nd relapse, 3rd relapse and transformation represented 2.5%, 2.1% and 1.8% of the population, respectively. For patient 6, negative clones represent 1.6%, 2.1% and 2.7% of the subclone population in FL diagnosis, tFL diagnosis and tFL relapse events. N-gly site-negative clones were found to arise from either a positive or a negative clone, through a single nucleotide variant. Several negative clones can arise from a shared positive ancestor, as depicted through the wide branching. Further SHM in these negative clones does not result in site re-acquirement or gain of new sites.

Discussion

The high propensity of relapse in FL patients suggests that current therapies are not successfully targeting the early aberrations needed to propagate disease, leading to acquirement of further mutations that reduce effective treatment options. Therefore, uncovering and targeting features of FL ancestral cells may offer durable outcomes for patients.

We report for the first time the behaviour of N-gly sites during disease progression by analysing the clonal repertoire of temporal FL samples based on *IGHV* sequencing. Samples ranged from diagnosis to transformation and included a mixed patient cohort with variable clinical disease courses, reflecting the heterogeneous nature of the disease (Supplemental Table 3). All patients harboured at least one acquired site in their earliest disease event that was conserved in both the heterogeneous subclonal population and the overall tumour mass. N-gly sites were also retained in sequential relapse and transformation samples although for patients 1 and 2, sites were conserved through non-synonymous mutations (Table 2). Analysis of the nine base pair region encoding the N-gly site for each patient revealed a group of subclones harbouring a different nucleotide sequence in the site to that of the major clone, due to ongoing SHM. However, the majority of these affected subclones maintained the N-gly site for patients 1, 2 (sites 1-3), 3, 4 (site 1) and 5 across disease events through synonymous and non-synonymous mutations (Figure 1a). Although the acquirement of additional 'driver' mutations through natural or therapy-related selection pressures may dampen the tumour's microenvironment dependency at later stages of disease, the conservation of N-gly sites suggest they retain an important functional significance**.**

The presence of negative subclones is an expected occurrence, as SHM does not differentiate between seemingly favourable and non-favourable mutations. Lineage trees have revealed how negative clones are derived from positive clones, suggesting acquirement of sites is an early event. As these negative clones represent only a small % of the tumour population, they are likely to be outcompeted by N-gly site-positive clones perhaps due to loss of the microenvironmental interaction provided via the added mannoses. Negative clones can still undergo SHM but do not reacquire sites in their progeny and, with the exception of Patient 3, are lost from subsequent samples, indicating that they are not selected to undergo expansion or long-term survival. Sanger sequencing of the light chain variable region for patient 3 did not reveal additional N-gly sites in the major clone. However, while we cannot assume that sites in the light chain are not acquired subclonally and may therefore be present in the negative subclones of patient 3, determining the light chain N-gly site status of *IGHV*-based subclones is currently impossible.

Patient 5 provided an interesting case for two reasons; the different anatomical sites for the two events and the discordant SHM within the *IGHV* between the two clonal populations. This discordance suggests an early divergence, where an ancestral cell with limited SHM, migrated from one site to another where selection pressures drove the outward growth of subclones with a specific SHM pattern. However despite *IGHV* sequence heterogeneity, the acquirement of N-gly sites within each population at different locations illustrates that sites are an essential feature of FL. The sharing of two subclones with the N-gly site of the diagnostic sample highlights the trafficking ability of site-positive subclones between anatomical sites, which is also observed in patient 6 (Figure 3). However for patient 6, SHM patterns between events were highly similar and the CDR3 N-gly site was conserved throughout, suggesting a late divergency between events from a shared ancestral cell**.** Therefore, N-gly sites are required in both early and late divergency models of evolution.

The conservation of N-gly sites within and across disease events and the lack of accumulation during ongoing SHM suggests they are an early and stable event in FL pathogenesis. Early events are usually determined through their conservation within temporal samples and for patients 1-3, WGS/WES had previously identified key genetic aberrations within a putative ancestral cell, known as the common progenitor cell (CPC). [19-21] Patients 1 and 3 had a 'sparse' CPC due to the lack of shared genetic aberrations across temporal samples, suggesting an early divergency with episodes arising from more genetically independent pathways. However, despite this mutational heterogeneity between events, N-gly sites are conserved, identifying an important feature of the CPC. This is a significant finding as

the CPC is believed to be the reservoir pool from which successive disease events arise, accounting for the high relapse rates experienced by the majority of patients. The latency between biopsy sampling (Supplemental Table 3), suggests an N-gly site-positive CPC that is able to remain dormant for many years before a mutational event leads to a new disease episode. The mannose-lectin interaction enables tumour retention and survival of the CPC permitting the accumulation of genetic events that lead to overt disease, suggesting a critical priming event in FL manifestation. Although epigenetic deregulation is a considered CPC event as evidenced in the previous genetic profiling of patient's 1-3 samples, our data implies that it is not solely sufficient for 'driving' the disease. Instead, it seems that the N-gly site profile determines which clones are able to expand and survive during disease progression, irrespective of the genetic profile of the subclones. Analysing the genetic profile of N-gly site-negative subclones will determine the validity of this hypothesis.

Although the t14:18 translocation can be found in healthy circulating B cells which do not go on to become malignant, [22-24] N-gly sites in the variable region are restricted to GC-derived lymphomas [3] indicating an attractive and tumour specific therapeutic target which may lead to the loss of a critical CPC-microenvironmental interaction and reduce the frequency of relapse. The presence of N-gly sites in the presumed FL precursor, ISFN[15] supports the theory of N-gly sites occurring at an early stage of pathogenesis, being acquired even before disease manifestation. Figure 7 summarises how N-gly sites impact the evolution of disease.

Acknowledgements

This work was supported by grants from the Pathological Society of Great Britain and Ireland, Leukaemia UK Charity, Barts Cancer Charity and The Greg Wolf Fund. The authors thank Genewiz (New Jersey, US) for performing next generation sequencing. The authors acknowledge the Tissue Bank at Barts Cancer Institute (UK) for providing patient samples and corresponding clinical information.

Authorship contributions

Contribution: M.O., E.C., S.A., J.O., F.S., M.C. and S.K. performed research and analysed data; M.O., J.G., F.K.S., F.F., M.C., and S.K. designed the research and analysed the data; S.A., J.O. provided patient samples and analysed clinical data; M.O. and S.K. wrote the initial draft of the manuscript. All authors contributed to the modification of the draft and approved the final submission.

Conflict of interest disclosures

The authors declare no competing financial interests.

References

1. Cleary, M.L. and J. Sklar, Nucleotide sequence of a t(14;18) chromosomal breakpoint in follicular lymphoma and demonstration of a breakpoint-cluster region near a transcriptionally active locus on chromosome 18. *Proc Natl Acad Sci U S A*, 1985. 82(21): p. 7439-43.

2. McCann, K.J., et al., Remarkable selective glycosylation of the immunoglobulin variable region in follicular lymphoma. *Mol Immunol*, 2008. 45(6): p. 1567-72.

3. Zhu, D., et al., Acquisition of potential N-glycosylation sites in the immunoglobulin variable region by somatic mutation is a distinctive feature of follicular lymphoma. *Blood*, 2002. 99(7): p. 2562-8.

4. Zabalegui, N., et al., Acquired potential N-glycosylation sites within the tumor-specific immunoglobulin heavy chains of B-cell malignancies. *Haematologica*, 2004. 89(5): p. 541-6.

5. McCann, K.J., et al., Universal N-glycosylation sites introduced into the B-cell receptor of follicular lymphoma by somatic mutation: a second tumorigenic event? *Leukemia*, 2006. 20(3): p. 530-4.

6. Kuppers, R. and F.K. Stevenson, Critical influences on the pathogenesis of follicular lymphoma. *Blood*, 2018. 131(21): p. 2297-2306.

7. Alcoceba, M., et al., Preferential Acquision of N-Glycosylation Sites in the VDJ Region in Germinal Center B-Cell-Like Difusse Large B-Cell Lymphoma. *Blood*, 2012. 120(21): p. 1589-1589.

8. Amin, R., et al., DC-SIGN-expressing macrophages trigger activation of mannosylated IgM B-cell receptor in follicular lymphoma. *Blood*, 2015. 126(16): p. 1911-20.

9. Linley, A., et al., Lectin binding to surface Ig variable regions provides a universal persistent activating signal for follicular lymphoma cells. *Blood*, 2015. 126(16): p. 1902-10.

10. Coelho, V., et al., Glycosylation of surface Ig creates a functional bridge between human follicular lymphoma and microenvironmental lectins. *Proc Natl Acad Sci U S A*, 2010. 107(43): p. 18587-92.

11. Strout, M.P., Sugar-coated signaling in follicular lymphoma. *Blood*, 2015. 126(16): p. 1871-2.

12. Schneider, D., et al., Lectins from opportunistic bacteria interact with acquired variable-region glycans of surface immunoglobulin in follicular lymphoma. *Blood*, 2015. 125(21): p. 3287-96.

13. Okosun, J., et al., Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat Genet*, 2014. 46(2): p. 176-181.

14. Green, M.R., et al., Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proc Natl Acad Sci U S A*, 2015. 112(10): p. E1116-25.

15. Mamessier, E., et al., Contiguous follicular lymphoma and follicular lymphoma in situ harboring N-glycosylated sites. *Haematologica*, 2015. 100(4): p. e155-7.

16. Carlotti, E., et al., High Throughput Sequencing Analysis of the Immunoglobulin Heavy Chain Gene from Flow-Sorted B Cell Sub-Populations Define the Dynamics of Follicular Lymphoma Clonal Evolution. *PLoS One*, 2015. 10(9): p. e0134833.

17. Jiang, Y., et al., Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. *Genome Biol*, 2014. 15(8): p. 432.

18. Barak, M., et al., IgTree: creating Immunoglobulin variable region gene lineage trees. *J Immunol Methods*, 2008. 338(1-2): p. 67-74.

19. Carlotti, E., et al., Transformation of follicular lymphoma to diffuse large B-cell lymphoma may occur by divergent evolution from a common progenitor cell or by direct evolution from the follicular lymphoma clone. *Blood*, 2009. 113(15): p. 3553-7.

20. Pasqualucci, L., et al., Genetics of follicular lymphoma transformation. *Cell Rep*, 2014. 6(1): p. 130-40.

21. Green, M.R., et al., Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood*, 2013. 121(9): p. 1604-11.

22. Limpens, J., et al., Lymphoma-associated translocation t(14;18) in blood B cells of normal individuals. *Blood*, 1995. 85(9): p. 2528-36.

23. Dolken, G., et al., BCL-2/JH rearrangements in circulating B cells of healthy blood donors and patients with nonmalignant diseases. *J Clin Oncol*, 1996. 14(4): p. 1333-44.

24. Schuler, F., et al., Prevalence and frequency of circulating t(14;18)-MBR translocation carrying cells in healthy individuals. *Int J Cancer*, 2009. 124(4): p. 958-63.

Tables

Patient	Disease Event	Dominant *VDJ* rearrangement	No of unique *VDJ* rearrangements	Dominant rearrangement (% of total reads)	No of unique subclones with dominant *VDJ* rearrangement
1	Diagnosis	V3-30, D3-16, J6	29	99.03	1690
	Transformation	V3-30, D3-16, J6	313	94.95	2727
2	1st relapse	V3-11, D3-16, J1	105	95.04	3530
	3rd relapse	V3-11, D3-16, J1	12	99.33	3032
	Transformation	V3-11, D3-16, J1	129	99.23	3966
3	2nd relapse	V3-48, D1-26, J4	164	88.23	4103
	3rd relapse	V3-48, D1-26, J4	261	71.10	2057
	Transformation	V3-48, D1-26, J4	143	84.81	1592
4	1st relapse	V3-48, D3-10, J6	101	56.60	81
	2nd relapse	V3-48, D3-10, J6	324	69.70	248
5	1st relapse	V3-23, D4-23, J6	224	53.30	140
	3rd relapse	V3-23, D3-3*, J5*	433	70.70	371
6	FL diagnosis	V3-23, D5-18, J6	7,931	82.40	2510
	tFL diagnosis	V3-23, D5-18, J6	2,436	89.60	3442
	tFL relapse	V3-23, D5-18, J6	1,204	93.10	4749

Table 1. Summary of *VDJ* sequencing results for patients 1-6. For patients 4 and 5, information in table was obtained from reference 16 and patient 6 information was obtained from reference 17. * Although samples of patient 5 have different *DJ* rearrangements according to IMGT, when aligned they show highly similar CDR3 regions and share a t(14:18) breakpoint, indicating a clonal relationship

Patient	Disease Event	No of N-gly sites	Region in *IGHV*	AA position of N-gly site	N-gly motif in major clone	% of unique subclones with N-gly site of major clone present
1	Diagnosis	1	CDR3	108	NFS	98.31
	Transformation	1	CDR3	108	***NVS***	98.17
2	1st relapse	4	CDR1, FR2, CDR2, FR3	30, 39, 56, 85	NFS, NMS, NIT, NNS	99.01, 99.01, 98.87, 97.22
	3rd relapse	4	CDR1, FR2, CDR2, FR3	30, 39, 56, 85	NFS, NMS, NIT, NNS	99.31, 96.8, 98.75, 81.83
	Transformation	4	CDR1, FR2, CDR2, FR3	30, 39, 56, 85	NFS, NMS, NIT, ***NNT***	98.54, 98.16, 97.71, 84.95
3	2nd relapse	1	CDR2	56	NIS	97.47
	3rd relapse	1	CDR2	56	NIS	97.91
	Transformation	1	CDR2	56	NIS	98.18
4	1st relapse	2	FR2, CDR3	48, 108	NKS, NNS	97.53, 100
	2nd relapse	2	FR2, CDR3	48, 108	NKS, NNS	97.98, 99.6
5	1st relapse	1	CDR1	30	NFS	97.14
	3rd relapse	1	FR3	94	***NLT***	99.19
6	FL diagnosis	1	CDR3	108	NGS	98.41
	tFL diagnosis	1	CDR3	108	NGS	97.94
	tFL relapse	1	CDR3	108	NGS	96.74

Table 2. N-gly sites identified in the major clone of six FL patients taken at different time points of disease**.** Italic text in red refers to differences in the location or amino acid (aa) sequence of N-gly sites, referred to as N-gly motif, within the major clone across temporal samples. The major clone was determined by the highest count number. The multiple values in patient 2 and 4 relate to the multiple N-gly sites observed. The majority of subclones across patients and disease events retain the N-gly site of the major clones (P<0.0001). AA; amino acid. Numbers given under the heading 'AA position of N-gly site' refer to the position of the middle amino acid making up the N-gly site according to IMGT numeration of the variable region.

Patient	Disease Event	No of unique tumour related subclones	No of unique tumour subclones with different codon sequence in N-gly site region to the MC (% of total subclones)	Subclones with different codon sequence in N-gly region	
				% of subclones without N-gly site	% of subclones with N-gly site
1	Diagnosis	2727	107 (3.92)	1.69	2.24
	Transformation	1690	110 (6.51)	1.83	4.67
2	1st relapse	3530	96 (2.72), 113 (3.20), 2062 (58.41), 226 (6.40)	0.99, 0.99, 1.13, 2.78	1.73, 2.21, 57.28, 3.63
	3rd relapse	3032	145 (4.78), 239 (7.88), 982 (32.39), 737 (24.31)	0.69, 3.10, 1.25, 18.17	4.09, 4.68, 31.13, 6.13
	Transformation	3966	134 (3.38), 148 (3.73), 1370 (34.54), 747 (18.84)	1.46, 1.84, 2.29, 15.05	1.92, 1.89, 32.25, 3.78
3	2nd relapse	4103	385 (9.38)	2.53	6.85
	3rd relapse	2057	179 (8.70)	2.09	6.61
	Transformation	1592	455 (28.58)	1.82	26.76
4	1st relapse	81	28 (34.57), 0	2.47, 0	32.1, 0
	2nd relapse	248	14 (5.65), 1 (0.40)	2.02, 0.40	3.63, 0
5	1st relapse	140	22 (15.71)	2.86	12.86
	3rd relapse	371	21 (5.66)	0.81	4.85
6	FL diagnosis	2510	946 (37.69)	1.59	36.10
	tFL diagnosis	3442	265 (7.70)	2.06	5.64
	tFL relapse	4749	214 (4.51)	2.70	1.81

Table 3. No. of unique subclones that have a different codon sequence at the location of the N-gly site to that of the major clone. The percentage of these subclones that are either motif positive or motif negative are provided. Percentages were calculated from the total number of unique subclones identified. Values were rounded to 2 decimal points.

	Patient 1		Patient 2			Patient 3			Patient 4		Patient 5		Patient 6		
Least mutated subclone	FL	tFL	FL1	FL3	tFL	FL2	FL3	tFL	FL1	FL2	FL1	FL3	FL	tFL	tFL1
% homology to G.L V gene sequence	96.4	98.0	88.7	88.3	94.4	84.7	86.1	95.6	89.4	89.0	98.0	89.1	97.8	93.0	92.3
N-gly site presence (Y/N)	Y	Y	Y, Y, Y, Y	Y, Y, Y, N	N, N, N, Y	Y	Y	N	Y,Y	Y, Y	N	Y (FL1 site)	Y	Y	Y
Most mutated subclone															
% homology to G.L V gene sequence	84.3	83.9	84.7	84.3	84.3	79.0	79.4	79.6	87.4	86.9	87.5	81.5	79.4	71.3	73.5
N-gly site presence (Y/N)	Y	Y	Y, Y, Y, Y	Y, Y, Y, Y	Y, Y, Y, Y	Y	Y	Y	Y,Y	Y,Y	Y	Y	Y	Y	Y
% difference in homology between least and most mutated subclones	12.1	14.1	4.0	4.0	10.1	5.7	6.7	16.0	2.0	2.1	10.5	7.6	18.4	21.7	18.8

Table 4: N-gly site status in most and least diverse subclone based on degree of SHM compared to *V* gene germline sequence. Sequences were analysed for presence or absence of the N-gly motif site (s) found in the major clone. FL-FL diagnosis, FL1-1st relapse, FL2-2nd relapse, FL3-3rd relapse, tFL-transformation tFL1-relapsed tFL.

Figure Legends

Figure 1. Subclones with distinct codon sequences in the 9 basepair region of the N-gly site region. a) N-gly site status of subclones with a different codon sequence in N-gly site region to that of the major clone. N-gly site positive subclones represented here do not display either the same asparagine or serine/threonine encoding nucleotide sequence as that of the major clone. The middle codon for all sequences was checked to ensure for the absence of proline as this middle amino acid would negatively affect the functionality of the N-gly site. Percentages were calculated from the values in columns 5 and 6 from Table 3. b) Patient 3 subclones from the three disease events are illustrated as an example. The pie charts provide an overview of the distribution of subclones harbouring either synonymous (green) or non-synonymous (orange) mutations compared to the sequence of the major clone. The numbers inside the pie charts are representative of the actual number of subclones. For the 2nd relapse and transformation subclones, the majority of subclones have a different amino acid sequence encoding for the N-gly site (non-synonymous) whereas the majority of the 3rd relapse subclones maintain the amino acid sequence found in the major clone, which is NIS (synonymous). The tables highlight the variety of amino acid sequences that encode for N-gly sites (e.g. NIT, NVT, NIS) found in these subclones and the diverse range of codon sequences that are responsible for both these synonymous and non-synonymous mutations. Numbers in brackets represent the number of unique subclones presenting with the particular codon sequence.

Figure 2. Early and late divergent evolution and N-gly acquirement. a) Representative of patient 5. N-gly sites are acquired in site-specific precursor cells, explaining the different N-gly sites we observe in the 1st and 3rd relapse events taken from distinct anatomical sites (Table 2). b) Representative of patient 6. Here, the N-gly site was acquired in a shared precursor cell before divergence of tumour populations to distinct anatomical sites. G.L is an abbreviation of germline. Letters in cells represent the N-gly site amino acid sequences.

Figure 3. Venn diagrams showing the number of shared and distinct subclones across disease events. Diagrams were generated by comparing the VDJ sequences of subclones in each disease event of a patient (Table 1) to identify the number of shared and unshared subclones. The numbers in the overlaps represent the number of shared subclones between disease events whereas the numbers outside the overlap represent the number of unique subclones present for the particular disease event. Values in brackets represent the number of subclones as a % of the total number of subclones identified across all disease events. These values were rounded to 1 decimal point. For patient 3, the transformation sample was omitted from analysis due to the difference in CDR3 length in comparison to the 1st and 2nd relapse. FL-FL diagnosis, FL1-1st relapse, FL2-2nd relapse, FL3-3rd relapse, tFL1-relapsed tFL.

Figure 4. Lineage trees for patient 4 showing the N-gly site status of each subclone. The left panel represents the lineage tree of the 1st relapse event whereas the right panel represents part of the lineage tree of the 2nd relapse event. Each subclone is represented by a node, major clones are indicated in larger nodes. Nodes are split into two colours to represent the two N-gly sites found in the FR2 and CDR3 regions for patient 4 (Table 2). FR2 site is represented in left half and CDR3 site is represented in right half of the nodes. Nodes differing in colour to the major clone represent subclones with different N-gly site codon sequences. Black represents absence of the FR2 or CDR3 site. Major clones between the two disease events have a different codon sequence in the FR2 N-gly site, indicated by the difference in colour. Nodes in boxes represent shared subclones between events. White nodes represent subclones inferred to exist but not detected through 454 sequencing. Germline nodes are in grey at the top of the tree, marked G.L. Numbers on branches indicate >1 mutation

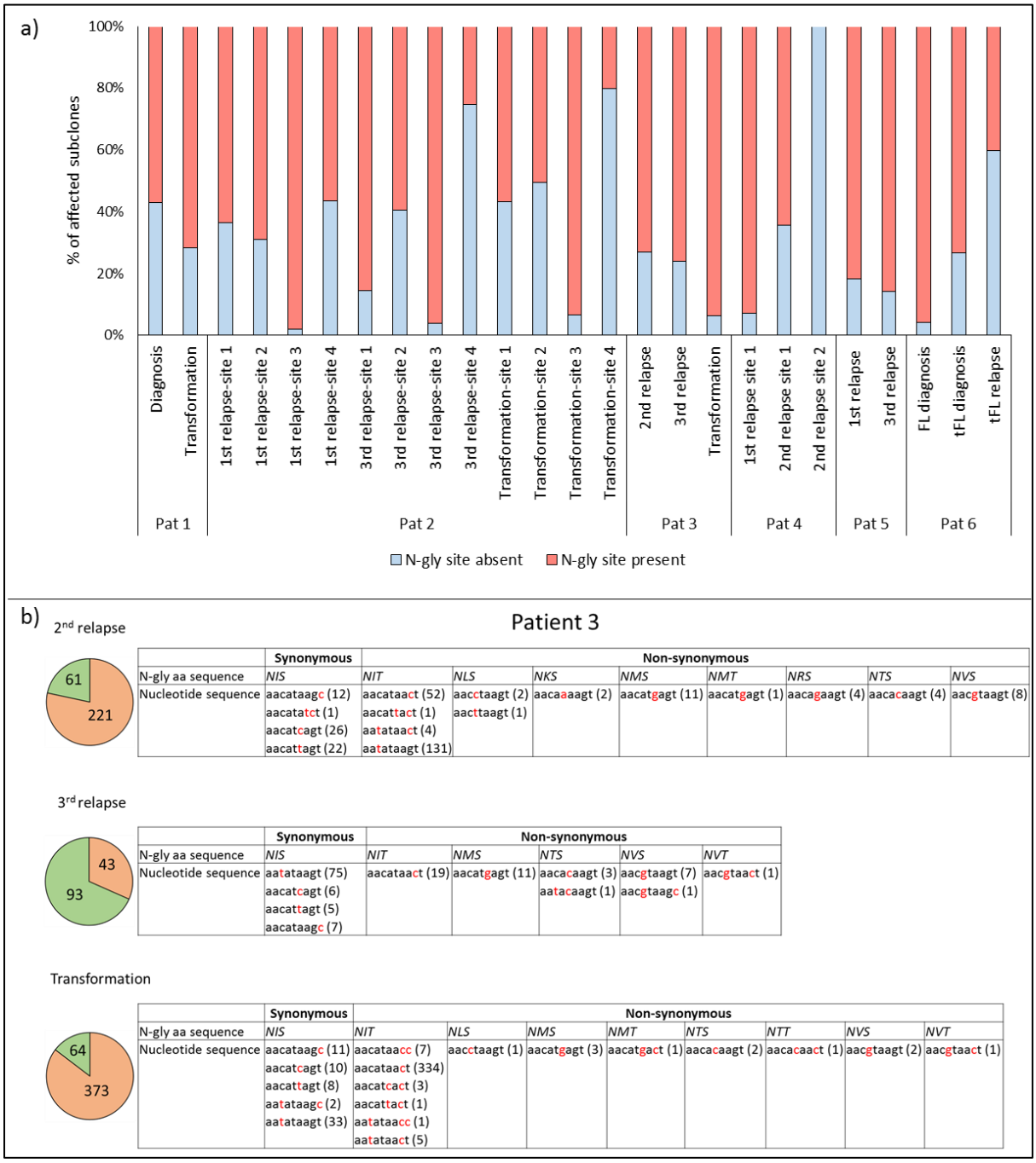
separating one node from the other. The full lineage tree of 2nd relapse event are provided in Supplemental Figure 2a. MC; major clone.

Figure 5. Lineage trees for patient 5 showing the N-gly site status of each subclone. The left panel represents the lineage tree of the 1st relapse event whereas the right panel represents part of the lineage tree of the 3rd relapse event. Black nodes represent nodes that are absent of N-gly site found in the major clone. Nodes differing in colour to the major clone represent subclones with different N-gly site codon sequences. The 4 nodes split into two colours (black and orange) in the right panel represent subclones which are absent for the N-gly site found in the major clone of the 3rd relapse (NLT) event but are positive for the N-gly site found in the 1st relapse event (NFS). Shared subclones are highlighted in boxes. The full lineage tree of the 3rd relapse event are provided in Supplemental Figure 2b.

Figure 6. Hierarchy of N-gly site-negative clones in relation to their direct ancestral clone for Patient 1, 3 and 6. This diagram represents part of the lineage trees as the majority of subclones (which are positive) are not shown and the interconnection between clones from a germline sequence is not depicted. Only the negative subclones, their parent clone and their progenitor clones are depicted here to highlight their generation and contribution to the clonal repertoire within an event. For all patients, black nodes represent N-gly site-negative clones. White nodes represent clones not detected by the sequencing platform but predicted to exist by the IgTree program. All white nodes are assumed to contain the N-gly site due to a number of other experimentally detected progeny clones being site-positive. Grey nodes in the patient 1 diagnosis event represent N-gly site positive subclones that contain the same N-gly site-encoding codon sequence to the major clone. Dark red nodes in patient 1 transformation represent N-gly site positive subclones that contain the same N-gly site-encoding codon sequence to the major clone of the event. Pink nodes in patient 3 represent N-gly site positive subclones that contain the same N-gly site-encoding codon sequence to the major clone of the event. Green nodes in patient 6 represent N-gly site positive subclones that contain the same N-gly site-encoding codon sequence to the major clone of the event. Other coloured nodes in trees represent N-gly site-positive clones which contains a different codon sequence in the N-gly site compared to the major clone of the disease event. The full lineage trees can be requested from the corresponding author.

Figure 7. Simplified model of FL evolution and progression based on the high throughput sequencing of the immunoglobulin heavy chain variable gene and N-gly analysis. Following the t14;18 translocation during a likely error in VDJ recombination in the bone marrow, the B cell migrates to the germinal centre where it undergoes SHM. N-gly sites are acquired early on in the process (purple figure), in the presumed precursor lesion, in situ follicular neoplasia (ISFN). As FL-like B cells are believed to represent the circulating counterparts of ISFN, N-gly sites may be retained in these cells also. Precursor cells which do not acquire sites through SHM undergo clonal deletion, assumed by the low frequency of site-negative subclones in the clonal repertoire. Clones maintain conservation of N-gly sites during tumour evolution through ongoing SHM and gain of additional mutations, leading to an ancestral cell pool population, the CPC. With the exception of patient 5, the CPC provides a reservoir from which distinct disease events arise from, which retain the N-gly site. At each stage of evolution, SHM results in the emergence of clones which lose N-gly sites due to the largely random nature of the process that does not distinguish between favourable and non-favourable mutations. However these represent only a minor mass of the heterogeneous tumour population and with the exception of patient 3, cannot traffic between events, indicating their insignificance in propagating progression and their likely loss from the clonal repertoire through cell death pathways. PFL - partial involvement by FL.

Figures



	Synonymous	Non-synonymous							
N-gly aa sequence	*NIS*	*NIT*	*NLS*	*NKS*	*NMS*	*NMT*	*NRS*	*NTS*	*NVS*
Nucleotide sequence	aacataagc (12) aacatatct (1) aacatcagt (26) aacattagt (22)	aacataact (52) aacattact (1) aatataact (4) aatataagt (131)	aacctaagt (2) aacttaagt (1)	aacaaaagt (2)	aacatgagt (11)	aacatgagt (1)	aacagaagt (4)	aacacaagt (4)	aacgtaagt (8)

	Synonymous	Non-synonymous				
N-gly aa sequence	*NIS*	*NIT*	*NMS*	*NTS*	*NVS*	*NVT*
Nucleotide sequence	aatataagt (75) aacatcagt (6) aacattagt (5) aacataagc (7)	aacataact (19)	aacatgagt (11)	aacacaagt (3) aatacaagt (1)	aacgtaagt (7) aacgtaagc (1)	aacgtaact (1)

	Synonymous	Non-synonymous						
N-gly aa sequence	*NIS*	*NIT*	*NLS*	*NMS*	*NMT*	*NTS*	*NTT*	*NVS*	*NVT*
Nucleotide sequence	aacataagc (11) aacatcagt (10) aacattagt (8) aatataagc (2) aatataagt (33)	aacataacc (7) aacataact (334) aacatcact (3) aacattact (1) aatataacc (1) aatataact (5)	aacctaagt (1)	aacatgagt (3)	aacatgact (1)	aacacaagt (2)	aacacaact (1)	aacgtaagt (2)	aacgtaact (1)

Figure 1. Subclones with distinct codon sequences in the 9 basepair region of the N-gly site region.

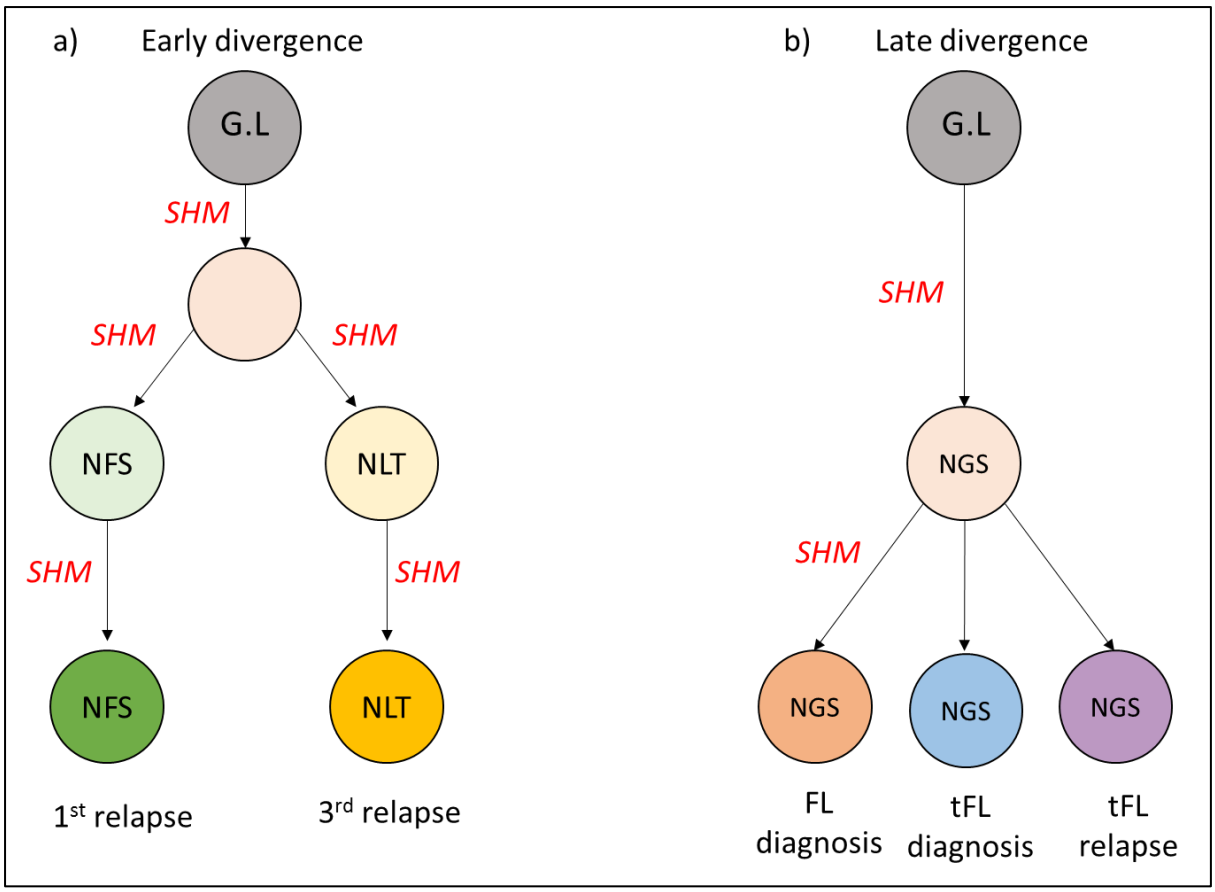


Figure 2. Early and late divergent evolution and N-gly acquirement.

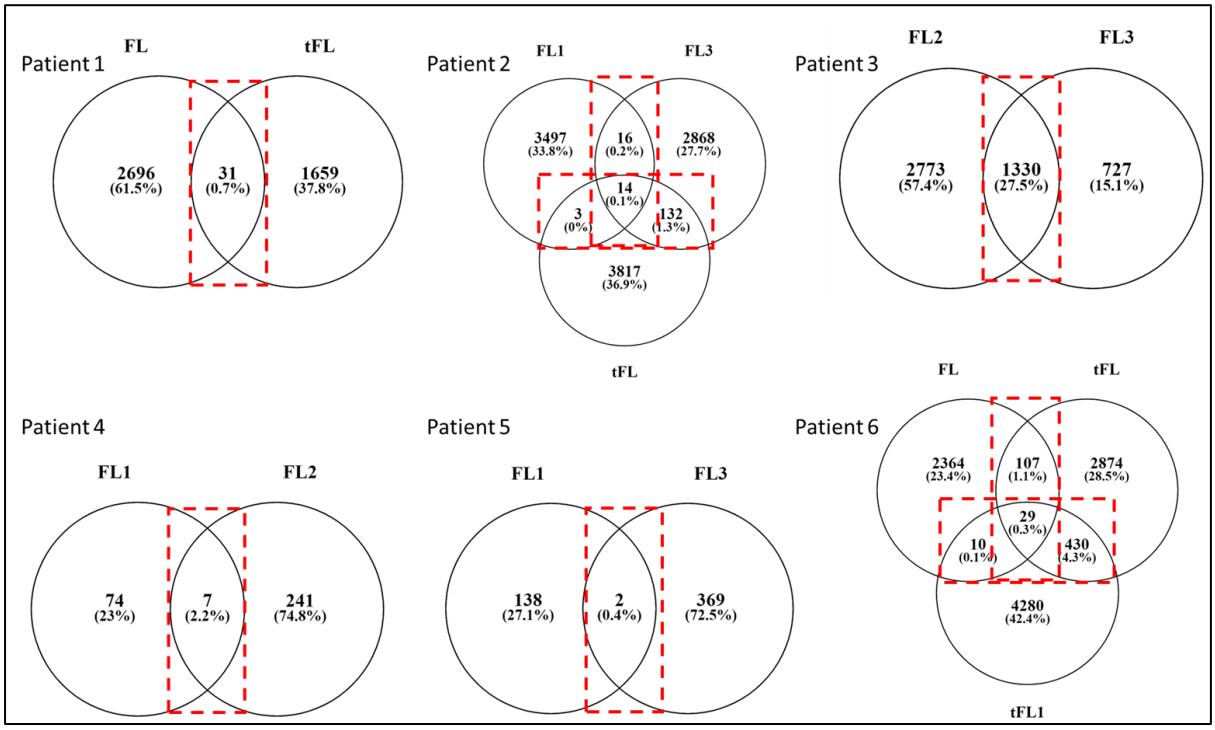


Figure 3. Venn diagrams showing the number of shared and distinct subclones across disease events.

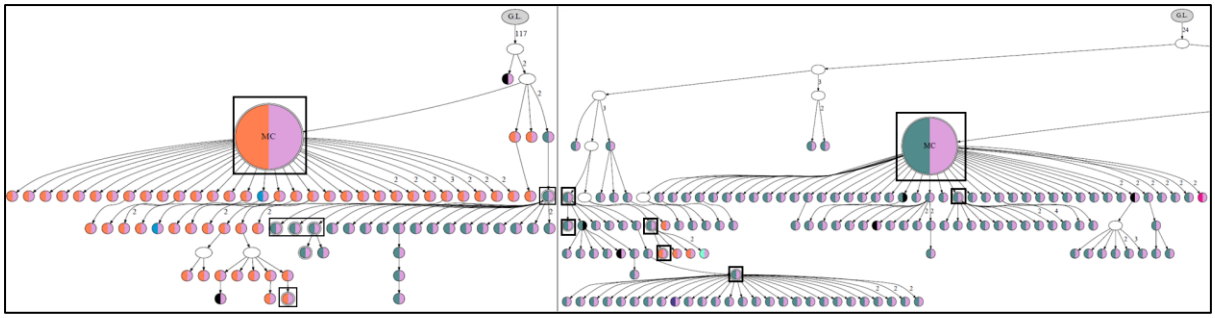


Figure 4. Lineage trees for patient 4 showing the N-gly site status of each subclone.

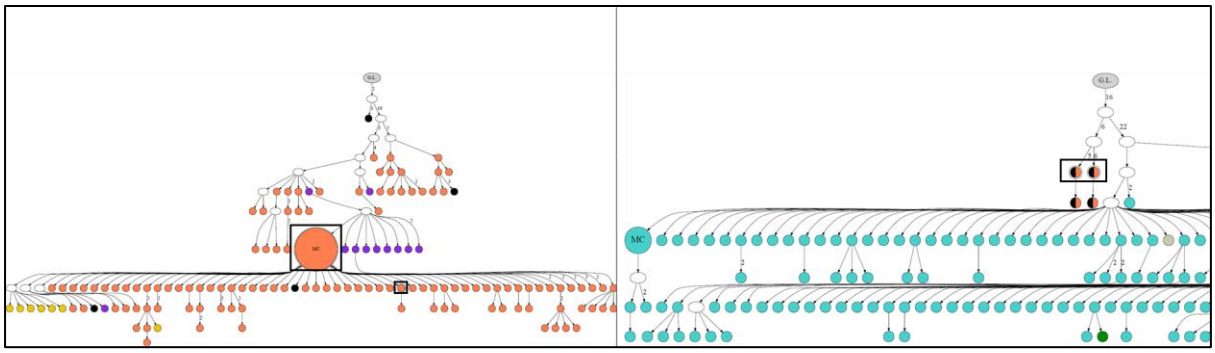


Figure 5. Lineage trees for patient 5 showing the N-gly site status of each subclone.

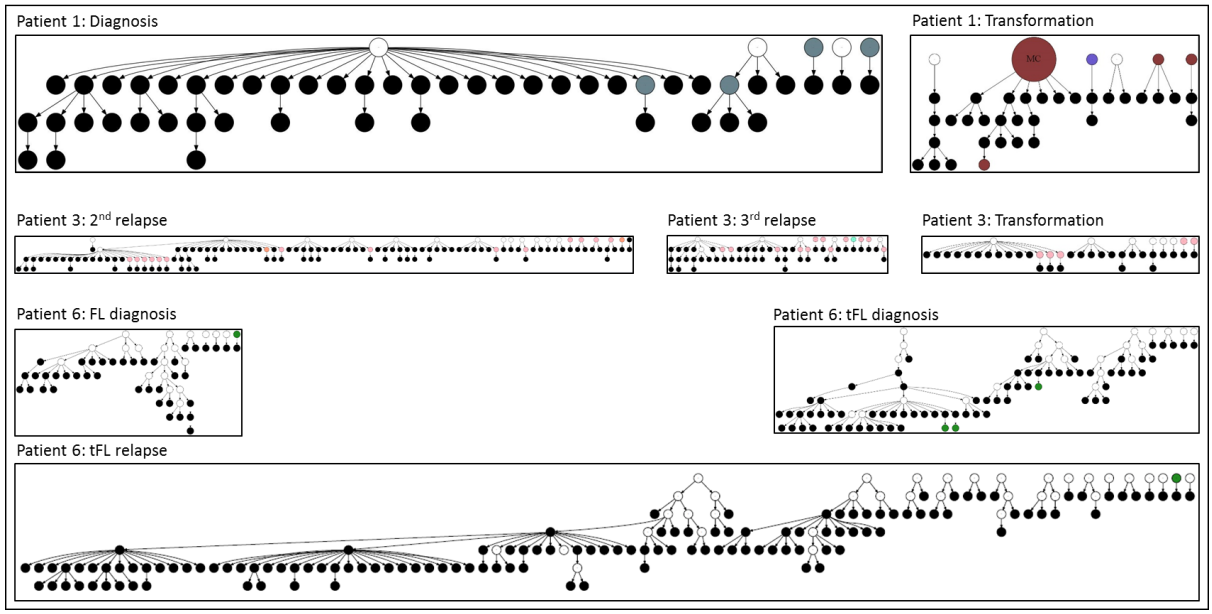


Figure 6. Hierarchy of N-gly site-negative clones in relation to their direct ancestral clone for patients 1, 3 and 6.

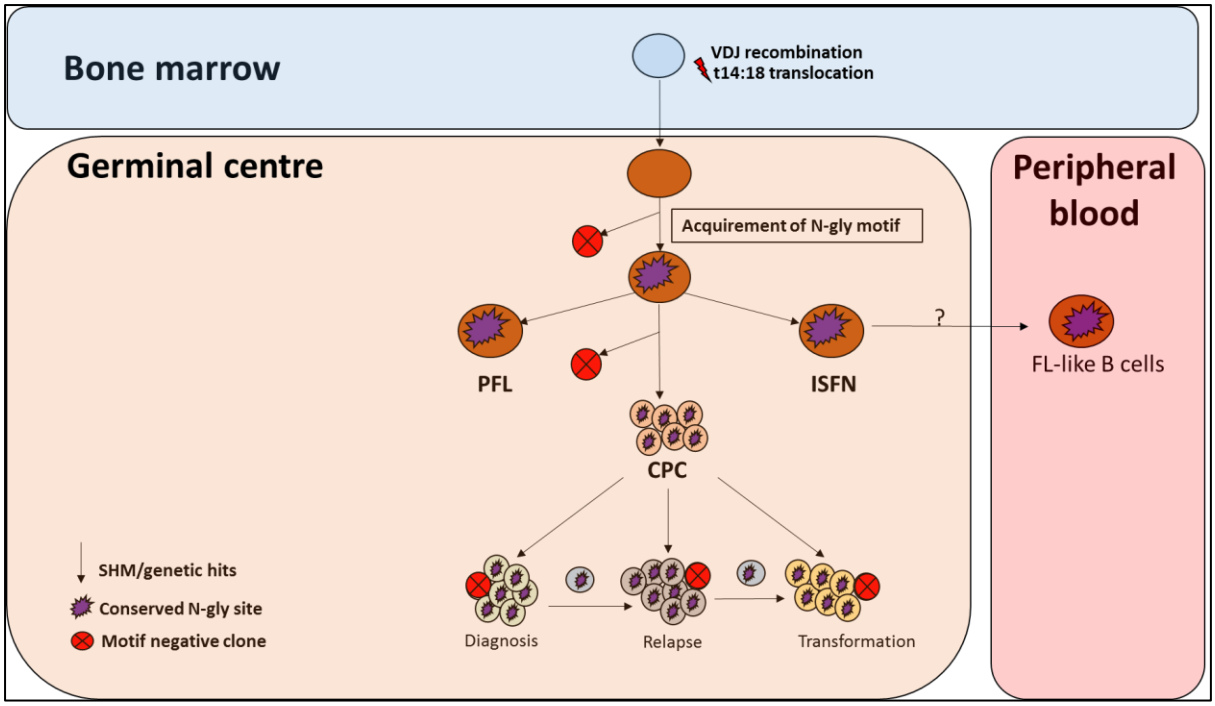


Figure 7. Simplified model of FL evolution and progression based on the high throughput sequencing of the immunoglobulin heavy chain variable gene and N-gly site analysis.