Immunoglobulin variable-region gene mutational lineage tree analysis: Application to autoimmune diseases

Avital Steiman-Shimony [1], Hanna Edelman [1], Michal Barak [1], Gitit Shahaf [1], Deborah Dunn-Walters [2], David I. Stott [3], Roshini Abraham[4], Ramit Mehr [1]

[1] Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, ISRAEL.

[2] Department of Immunobiology, King's College London, GKT Medical School, London, UK.

[3] Department of Immunology & Bacteriology, Division of Immunology, Infection and Inflammation, University of Glasgow, Western Infirmary, Glasgow G11 6NT, Scotland, U.K.

[4] Division of Hematology , Department of Internal Medicine and Division of Clinical Biochemistry and Immunology, Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, MN, USA.

Corresponding Author:

Dr. Ramit Mehr, Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, ISRAEL.

Tel: +972-3-531-7990, Fax: +972-3-535-1824, Email: mehrra@mail.biu.ac.il

Grant support: The work was supported in parts by grants from the Israel Science Foundation (grant number 759/01-1) and the Israel Cancer Research Fund (to RM); the BBSRC (to DDW); The Mayo Clinic Hematologic Malignancies Research Program (RSA); RM is also supported by grants from the Human Frontiers Science Program and the Swedish Foundation for Strategic Research (funding the Strategic Research Center for studies on Integrative Recognition in the Immune System, IRIS, Karolinska Institute, Stockholm, Sweden, where RM was on sabbatical during the writing of this article). DIS's research was supported by grants from The Wellcome Trust (054449/Z/98/Z), the European Commission (ERB 4001 GT 950115) and the Chief Scientist Office (K/MRS/50/c2738).

Abstract

Lineage trees have frequently been drawn to illustrate diversification, via somatic hypermutation (SHM), of immunoglobulin variable-region (*IGV*) genes. In order to extract more information from *IGV* sequences, we developed a novel mathematical method for analyzing the graphical properties of IgV gene lineage trees, allowing quantification of the differences between the dynamics of SHM and antigen-driven selection in different lymphoid tissues, species, and disease situations. Here, we investigated trees generated from published *IGV* sequence data from B cell clones participating in autoimmune responses in patients with Myasthenia Gravis (MG), Rheumatoid Arthritis (RA), and Sjögren's Syndrome (SS). At present, as no standards exist for cell sampling and sequence extraction methods, data obtained by different research groups from two studies of the same disease often vary considerably. Nevertheless, based on comparisons of data groups within individual studies, we show here that lineage trees from different individual patients are often similar and can be grouped together, as can trees from two different tissues in the same patient, and even from IgG- and IgA-expressing B cell clones. Additionally, lineage trees from most studies reflect the chronic character of autoimmune diseases.

Key words

B lymphocyte, somatic hypermutation, lineage tree, autoimmune disease, bioinformatics.

Take-home messages

- *IGV* lineage trees contain useful quantitative information on the dynamics of the humoral immune response under various conditions.
- Lineage trees from B cell clones participating in autoimmune diseases (MG, RA, and SS) are often larger (in terms of overall number of nodes) than those from normal human germinal centers (GCs), most probably due to the long-term ongoing *IGV* diversification in these clones.
- More detailed analyses are hindered by the fact that data obtained by independent research groups often differ greatly, due to variability in cell sampling and sequence extraction methods, and possibly also to differences in disease duration by the time of sampling, and other patient or treatment factors.
- Lineage trees from IgA- and IgG- expressing B cell clones from the same patient do not significantly differ. Neither do lineage trees from different tissues in the same patient, e.g. labial salivary glands and lymph node, or parotid gland and peripheral blood, in the same SS patients.

Introduction

The generation of “lineage trees" (also called “dendrograms” or “pedigrees”) to visualize the lineage relationships of B cell mutants in the GCs has been used in the past to confirm the role of the GC as the location of somatic hypermutation [1,2,3], to identify lineage relationships between cells from independent GCs [4] or different tissues [5,6] and from additional processes of diversification such as gene conversion in the rabbit [7]. The experimentally generated lineage trees reflect the multiple rounds of mutation for each germline *IGV* that participates in the primary response. We have suggested that much information about the dynamics of antigen-driven clonal selection during the immune response is contained in the shape of lineage trees deduced from the final responding clones [8]. For example, trees generated from clones during the peak of the primary response are much more branched or ``bushy" [9], but trees become less branched as the response progresses [10]. The “pruned" shape of these trees has been cited as evidence of the destructive nature of somatic hypermutation. Other examples of lineage trees drawn to illustrate various aspects of the germinal center reaction, or differences in this reaction under varying circumstances, abound in the literature. In the above-reviewed studies, however, lineage tree classification has been based only on a qualitative, intuitive assessment of the most obvious shape characteristics, looking only at a few trees at a time.

In order to extract the quantitative information embedded in the shape characteristics of lineage trees, we developed a rigorous computer-aided algorithm for measuring lineage tree graphical properties (Figure 1 and Table 1), and have found correlations between these measures and the dynamical parameters of the GC response that generated the trees [11,12]. We have applied this method to study age-related differences in the humoral immune response in humans, with the surprising result that the GC reaction dynamics – and the effects of aging on these dynamics – differ between GCs taken from different tissues [13, 14]. Applying this method to lineage tree data from Ig primary and secondary diversification processes in rabbits and chickens, which use gene conversion as well as SHM during both primary and secondary diversification [7], has highlighted the unique characteristics of the genetic and selective processes driving the formation of the Ig repertoire in these two species. In our most recent study [15], we have analyzed the bone marrow (BM) and peripheral blood (PB) clonal populations from patients with Light Chain Amyloidosis (AL) and multiple myeloma for evidence

of ongoing B cell diversification. We detected the presence of clonally-related populations within the BM and PB, indicative of ongoing mutation. The clonal trees for these populations differed from those of normal controls. Differences in tree shape characteristics have also been observed in our preliminary analyses of *IGV* from several B cell malignancies (Zuckerman *et al,* unpublished).

Quantitative analysis of the shape properties of *IGV* lineage trees thus provides novel insights into the mechanisms of normal and malignant B cell clonal evolution. Our aim here is to broaden the application scope of lineage tree analysis towards other situations where *IGV* are diversified. In the present paper we review the results of our comparative analyses of *IGV* lineage trees from several autoimmune diseases, in which GC-like areas exist within the afflicted tissue or organ, composed mostly of activated B cells undergoing somatic hypermutation and (presumably antigen-driven) selection [16, 17, 18]. It has been suggested [16] that perhaps the onset of tolerance breakdown leading to autoimmune pathology is a result of B cell expansion in GCs outside the secondary lymphoid structures. These ectopic GCs may lack the normal protective mechanisms preventing expansion of auto-reactive mutants, thereby allowing an anti-self reaction to persist.

We have analyzed data from studies on patients with Myasthenia Gravis (MG), Rheumatoid Arthritis (RA), and Sjögren's Syndrome (SS). Different research groups use different methods of cell sampling and of sequence data extraction, there are also differences in the disease state and treatment of patients. These differences cause difficulties in making comparisons across different data sets. Nevertheless, we were able to identify several similarities and differences between data from different tissues, B cell isotypes and disease groups, as discussed below.

Methods

Source of data for analysis. The data used in this study were collected by a thorough literature search for *IGV* sequences from patients with autoimmune disease, this yielded the published studies listed in Table 2.

Lineage tree generation. In some of the studies we used [17-19], lineage trees were already drawn. In others, only sequences were published [20-23]; in these cases, we identified germline genes by

alignment with published human *IGV* [24], and trees were generated using a computer program developed by us (M. Barak *et al,* unpublished), which is specifically tailored to deal with *IGV* sequences.

Tree shape analysis. A *lineage tree* is defined, graphically, as a rooted tree where the nodes correspond to B cell receptor gene sequences (Figure 1A). For two nodes *X* and *Y*, we say that *Y* is a child of *X* if the sequence corresponding to *Y* is a mutant of the sequence corresponding to *X*, which differs from *X* by only one mutation, and is one mutation further than *X* away from the original (germline) gene, that is, the root. Two B cells with identical receptor genes will thus correspond to the same node. A lineage tree depicts the maturation process of a B cell clone at a certain moment of observation – it consists only of the *IG* sequences of cells that were sampled at that moment and their ancestors back to the root, which were not necessarily sampled at the time of observation. Nodes in the tree (Figure 1A) can be either the *root* node, *leaves* (end-point sequences), or *internal nodes*, which can be either *split nodes* (branching points) or *pass-through nodes*.

The shape of lineage trees is quantified by measuring a number of properties of the graph describing the tree. For quite a few of these properties, the maximum, minimum and average values *per tree* may be measured. The complete list of variables measured is given in Table 1, including remarks on the meaning and usefulness of each variable. Our tree-measurement computer program reads a tree in the adjacency list format [11] and calculates the graphical variables. We intend to create, in the near future, a user-friendly interface for this program and make it available on the web for any researcher interested in analysis of their lineage trees; for the time being we are willing to analyze data sent to us upon request (see http://repertoire.ls.biu.ac.il/TREES for details).

Statistical analysis. Significance analysis was done using Student's t-test. To correct for the fact that we perform multiple comparisons (on 25 different tree properties), we used the FDR method [25]. Only differences that were found to be significant after this correction are considered as meaningful in our studies.

Results

Even when studying the same disease, different research groups use different experimental techniques (Table 2), and sometimes include different data groups within a study of particular disease (e.g. some of the IGV studied in RA were of different isotypes). Additionally, there may be differences caused by the method used for the generation of trees. There is more than one commonly used algorithm for generating phylogenetic trees from sequence data [26-29]. Different algorithms, or even different implementations of the same basic algorithm in different software packages, generate slightly different trees from the same data sets (Mehr et al, unpublished observations). It is not clear *a priory* whether these trees are statistically equivalent in their graphical properties. Hence we have made comparisons between all smallest possible groups in the first instance to determine which would be appropriate to combine.

We have used only one study of IgV gene diversification in MG, performed by Sims et al [17]. However, some lineage trees were generated by the authors [17] and the rest were generated by us from additional sequence data, so we first checked whether lineage trees in the two groups differ in any of the tree properties we measure. Fortunately, we found that the two MG data groups, containing 6 and 7 trees, respectively, do not significantly differ in tree properties, and hence may be grouped together.

The RA data was obtained from three separate studies done by different research groups, two using microdissection [19,20], and one using mononuclear cells isolated from single cell suspensions [23]. A comparison between tree properties from the first two groups yielded p-values smaller than 0.05 for only a few variables, and even those differences were insignificant after the FDR correction. In spite of this, we do not feel that the two data groups may be grouped together, as the average number of nodes (N) in the Gause et al.[20] trees is almost four-fold larger than that of the Kim et al.[19] trees (Figure 1B). Many other tree properties, such as the average root-to-leaf path lengths (average number of mutations per endpoint sequence), are proportionally larger in the former group's trees as well (Figure 1B). This difference could be due to different experimental methods, as each group's methods vary slightly at all stages of data extraction (sampling, cell labeling, and DNA amplification – see Table 2), or to different disease durations at time of extraction, and do not necessarily indicate an

inherent difference between B cell clone dynamics in these RA patients as would be expected if there were differences in disease severity. The first possibility definitely influences the results – although the pick size (number of PCR clones sequenced) in each experiment was not given (neither was disease duration in most cases), it is obvious that larger trees were observed wherever a large number of PCR clones was sequenced, as expected. In the data from Miura et al [23], IgA- and IgG-expressing B cells from two patients were analyzed separately; however some clones had cells from both isotypes, so we grouped them together and named these mixed trees "IgA-IgG". Upon measuring the trees, we found no significant differences between IgA-only, IgG-only, and IgA-IgG lineage trees, and hence these groups could be merged. Data from the two patients could also be merged as we did not find significant differences between the tree properties from the two patients. The single cell extraction method used in this study may have allowed for more cells to be extracted and more complete trees to be found; this is suggested by the larger number of leaves, L, and a result the number of nodes in these trees is also larger (Figure 1B). Pick size does not explain, however, why the number of mutations per sequence (average or max PL) is also so much larger in this data group. All in all, we cannot say whether the significantly larger trees in this data group, compared to the other two groups, mean that there is an intrinsic difference between B cell clone dynamics in the three studies, or are due to methodological differences.

A similar situation applies to the data from SS patients (Figure 1B), where trees in one microdissection study [21] were much larger – significantly larger this time – than the trees from a second microdissection study [18]. Within the data group in [21] there were sequences extracted from two tissues, salivary gland and lymph node. The salivary gland sequences were from two patients. The first comparison was thus made between these two patients; the differences between properties of trees generated from the sequences of these two patients were not significant, and hence the two data sets could be merged. We next compared tree properties in the two different tissues in [21]. The two tissue data sets showed no significant differences, and could therefore be also combined into one data set. However, when comparing the data from the two research groups [18,21], we found significant differences in the overall number of nodes and in 12 (out of 24 measured) other tree properties, all of which represent path lengths or parts of paths. A third study [22], which used mononuclear cells isolated from single cell suspensions, yielded lineage trees of intermediate sizes

compared to the above two groups. These data were obtained from two tissue sources, the parotid gland and peripheral blood. There was only one significant difference (in minimum DLFSN, see Table 1) between the two groups, yet the absolute values were similar, hence we treated these two groups as one. On the other hand, the maximum and average path lengths significantly differed between the Gellrich and Jacobi [21, 22] data; this could not be due to differences in sampling (which can change the numbers of leaves found and hence bushiness-related measures, but not the number of mutations in individual cells) and hence may be due to either a longer, or a more vigorous, proliferation in the Gellrich [23] clones.

Finally, we compared all the above data groups to data from normal human controls, taken either from microdissection of normal human GCs [13, 30] or from PBL [15]. What is immediately obvious from this comparison (Figure 1B) is that, in all but two of the data groups, lineage trees from autoimmune diseases are larger than those from normal immune responses. This is seen in spite of the fact that the control data could also come from clones participating in secondary responses. In the data groups which did not significantly differ from normal trees, the most likely reasons for this are that the relevant B cell clones in these cases had not yet had a chance to diversify significantly by the time of sampling. We believe that this difference is due to the chronic nature of autoimmune responses, and intend to investigate additional tree properties, to see if more can be learned about the dynamics of autoimmune GC reactions using this method.

Discussion

Lineage tree variables depend on the following four factors. The first important factor is the number of different sequences within each clone obtained in a given experiment – the “pick size”. This in turn is affected by the method of cell collection for DNA analysis, which also influences data resolution [12]. The number of distinct endpoint sequences (leaves) will depend on the pick size, and if the sample is small, this will also affect the tree’s degree of branching (“bushiness”). The experimental data analyzed above was generated by different research groups at different times, using various methods to extract DNA sequences from samples of patients in which the disease duration and severity probably varied as well. In order to extract meaningful conclusions from such data, it would help if

future studies use similar methods as much as possible, analyze a larger number of patients per study, and always include normal control groups as well. Moreover, the methods of tree generation from sequence alignments should also be standardized. Since generating IgV gene trees differs from generation of phylogenetic trees of other genes, e.g. evolutionary trees, where the root sequence is unknown, and, on the other hand, the trees can be assumed to be binary. Hence we have had to create our own program, specifically suited for IgV gene lineage tree generation (M. Barak *et al,* unpublished).

The second factor is the intrinsic potential for improvement via hypermutation possessed by each antigen receptor gene – the closer the original receptor is to optimally binding the driving antigen, the fewer the mutations that could lead to improvement of binding, and the higher the probability that a mutation will be deleterious or reduce the receptor's affinity to the antigen [8]. This was observed in a comparison of the response to a high affinity antigen (hen egg lysozyme) and a low affinity antigen (duck egg lysozyme) by B-cells expressing the same transgenic antigen receptor [31]. Affinity selection was only observed during the response to the low affinity antigen. This is mostly relevant in normal responses; in chronic responses, the GC reaction may go on even when the optimal receptor for the driving antigen has already been created.

The third factor or rather set of factors influencing lineage tree shape are the rate and duration of the hypermutation process itself. The overall number of mutations will depend on the product of the duration of the response (measured e.g. in days), the cells' probability of division per time unit (e.g. per day), and the probability of mutation per base pair per division. We assume that the hypermutation rate itself is already maximized, as it is already 10^6 times faster than normal somatic mutation, and is believed to occur by a single common mechanism; hence the overall number of nodes (individual mutations) in the tree, or the number of nodes from the root (the unmutated germline sequence) to a particular leaf (path length), are interpreted as reflecting the duration of the reaction and the cells' division rate. These two biological parameters probably exert the strongest influence on the data we have analyzed above, as demonstrated by the large variability in path lengths between groups.

Finally, tree shape is strongly influenced by the stringency of selection operating on the mutating cells, as selection “prunes” the trees, hence a tree’s bushiness is interpreted as an inverse measure of selection. We have yet to investigate this point in relation to autoimmune disease lineage trees. The relative strengths of the two opposing forces – diversification and selection – depend on the tissue in which the response occurs, the antigen, and other factors regulating response dynamics [12]. Since the absolute values of these rates cannot be directly measured, we study the dependence of tree properties on biological rates using computer simulations of humoral response dynamics [12, and ongoing work].

It should be noted that it is not known whether all clones in autoimmune ectopic GCs are specific for the self-antigen involved in the etiology of the autoimmune disease, or whether these environments are generally defective in regulation of B cell selection. In the latter instance there may be many irrelevant clones developing alongside the auto-reactive clones of interest. Only the MG data in [17] was generated from clones that were identified as being specific for the pathogenic self-antigen (acetylcholine receptor, AchR); hence, in the other cases reported here, the data may include both disease-specific and non-specific clones.

In summary, we find that lineage trees from B cell clones isolated from patients with autoimmune diseases (MG, RA, and SS) are often larger than those from normal human germinal centers (GCs), most probably due to the long-term ongoing *IGV* diversification in these clones. On the other hand, lineage trees from IgA- and IgG- expressing B cell clones from the same patient do not significantly differ. Neither do lineage trees from different tissues in the same patient, e.g. labial salivary glands and lymph node, or parotid gland and peripheral blood, in the same SS patients. More detailed analyses, which could supply much information about disease triggering and subsequent dynamics, are hindered by the fact that data obtained by independent research groups often differ greatly, due to variability in cell sampling and sequence extraction methods, and possibly also to differences in disease severity, and other patient or treatment factors.

Acknowledgements. The authors are indebted to R. Mage, D. Margolin, M. Shlomchik, F. Stevenson, and A. Brauninger, for sharing data and insights during the development of the tree analysis method, and to M. K. Manske and A. Sohni for generating the normal PBL data.

REFERENCES

[1] Kocks C, Rajewsky K. Stepwise Intraclonal maturation of antibody affinity through somatic hypermutation. Proc Natl Acad Sci 1988; 85: 8206-8210.

[2] Manser T. Evolution of antibody structure during the immune response. J Exp Med 1989;170:1211-1230.

[3] Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. Nature 1991;354:389-392.

[4] Vora KA, Tumas-Brundage K, Manser T. Contrasting the in situ behaviour of a memory B cell clone during primary and secondary immune responses. J Immunol 1999;163:4315-4327.

[5] Dunn-Walters DK, Boursier L, Ciclitira PJ, Spencer J. Immunoglobulin genes from human duodenal and colonic plasma cells are mutated. Biochem Soc Trans 1997;25:324S.

[6] Dunn-Walters DK, Isaacson PG, Spencer J. Sequence analysis of human IgVH genes indicates that ileal lamina propria plasma cells are derived from Peyer's patches. Eur J Immunol 1997;27:463-467.

[7] Mehr R, Edelman H, Sehgal D, Mage R. Analysis of Mutational Lineage Trees from Sites of Primary and Secondary Immunoglobulin Gene Diversification in Rabbits and Chickens. J Immunol 2004;172:4790-4796.

[8] Shannon M, Mehr R. Reconciling repertoire shift with affinity maturation: The role of deleterious mutations. J Immunol 1999;162:3950-3956.

[9] Jacob J, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. J Exp Med 1992;176:679-687.

[10] Jacob J, Przylepa J, Miller C, Kelsoe G. In situ studies of the primary immune response to (4-hydroxy-3- nitrophenyl)acetyl. III. The kinetics of V region mutation and selection in germinal center B cells. J Exp Med 1993;178:1293-1307.

[11] Dunn-Walters DK, Belelovsky A, Edelman H, Banerjee M, Mehr R. The Dynamics of Germinal Centre Selection as Measured by Graph-Theoretical Analysis of Mutational Lineage Trees. Dev Immunol 2002;9:233-245.

[12] Dunn-Walters DK, Edelman H, Mehr R. Immune System Learning and Memory Quantified by Graphical Analysis of B-Lymphocyte Phylogenetic Trees. BioSystems 2004;76:141-155.

[13] Banerjee M, Mehr R, Belelovsky A, Spencer J, Dunn-Walters DK. Age and tissue-specific differences in human germinal centre B cell selection. Eur J Immunol 2002;32:1947-1957.

[14] Dunn-Walters DK, Banerjee M, Mehr R. Age effects on antibody affinity maturation. Biochem Soc Trans 2003;31:447-448.

[15] Abraham RS, Manske MK, Sohni A, Edelman H, Zuckerman NS, Shahaf G, Dispenzieri A, Lacy MQ, Gertz MA, Mehr R. Analysis of B cell clonal evolution in BM and PBL of AL and MM patients using graphical quantification of lineage trees. In “Amyloid and Amyloidosis”. Eds.G. Grateau, R. A. Kyle, M. Skinner, CRC Press, Boca Raton. 2005:61-63.

[16] William J, Euler C, Christensen S, Shlomchik MJ. Evolution of Autoantibody Responses via Somatic Hypermutation Outside of Germinal Centers. Science 2002;297:2066-2070.

[17] Sims GP, Shiono H, Willcox N, Stott DI. Somatic Hypermutation and Selection of B Cells in Thymic Germinal Centers Responding to Acetylcholine Receptor in Myasthenia Gravis. J Immunol 2001;167:1935-1944.

[18] Stott DI, Hiepe F, Hummel M, Steinhauser G, Berek C. Sjögren’s Antigen-driven proliferation of B cells within the Target Tissue of an Autoimmune Disease-The Salivary Glands of Patients with Sjögren's Syndrome. J Clin Invest 1998;102(5):938-946.

[19] Kim HJ, Krenn V, Steinhauser G, Berek C. Plasma Cell Development in Synovial Germinal Centers in Patients with Rheumatoid and Reactive Arthritis. J Immunol 1999;162:3053-3062.

[20] Gause A, Gundlach K, Zdichavsky M, Jacobs G, Koch B, Hopf T, Pfreundschuh M. The B lymphocyte in rheumatoid arthritis: analysis of rearranged V kappa genes from B cells infiltrating the synovial membrane. Eur J Immunol 1995;25:2775-82. and data published in Genbank; Accession # X85162-X85167, Z75254, Z75333-Z75343, Z75442-Z75465.

[21] Gellrich S, Rutz S, Borkowski A, Golembowski S, Grominica-Ihle E, Sterry W, Jahn S. Analysis of VH-D-JH Gene transcripts in B cells infiltrating the salivary glands and lymph node tissues of patients with Sjögren's Syndrome. Arth Rhuem 1999;42:240-247.

[22] Jacobi AM, Hansen A, Kaufmann O, Pruss A, Burmester GR, Lipsky PE, Dorner T. Analysis of immunoglobulin light chain rearrangements in the salivary gland and blood of a patient with Sjögren's syndrome. Arthritis Res 2002;4:R4.

[23] Miura Y, Chu CC, Dines DM, Asnis SE, Furie RA, Chiorazzi N. Diversification of the Ig Variable Region Gene Repertoire of Synovial B Lymphocytes by Nucleotide Insertion and Deletion. Mol Med 2003;9:166-174.

[24] VBASE; http://vbase.mrc-cpe.cam.ac.uk/

[25] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B 1995;57:289-300.

[26] GeneWorks 2.5; http://www.med.nyu.edu/rcr/rcr/news/news_dec97.html#genewk

[27] Paup; http://www.accelrys.com/products/gcg_wisconsin_package/program_list.html

[28] GrowTree; http://www.accelrys.com/products/gcg_wisconsin_package/program_list.html

[29] Phylip; http://evolution.genetics.washington.edu/phylip.html

[30] Kuppers, R., Zhao, M., Hansmann, M.L., Rajewsky, K. Tracing B cell development in human germinal centres by molecular analysis of single cells picked from histological sections. EMBO J., 1993;12:4955-4967.

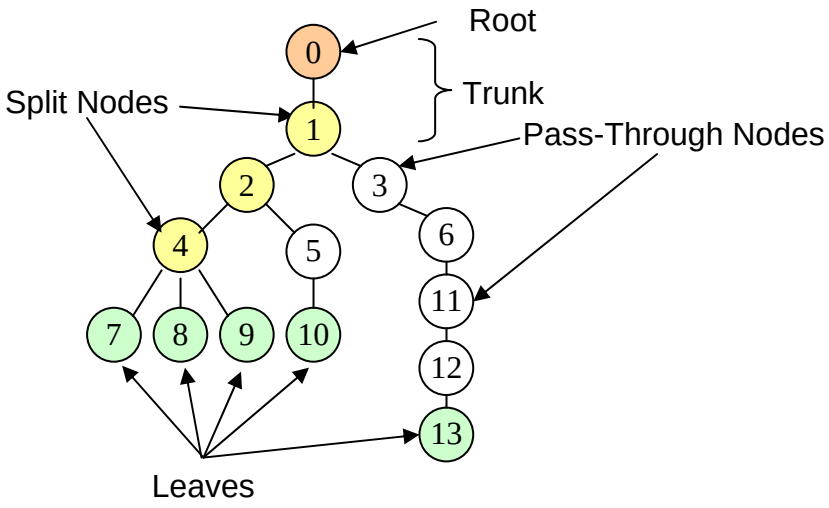
[31] Adams CL, Macleod MK, James Milner-White E, Aitken R, Garside P, Stott DI. Complete analysis of the B-cell response to a protein antigen, from in vivo germinal centre formation to 3-D modelling of affinity maturation. Immunology 2003;108:274-87.

Figure Legend

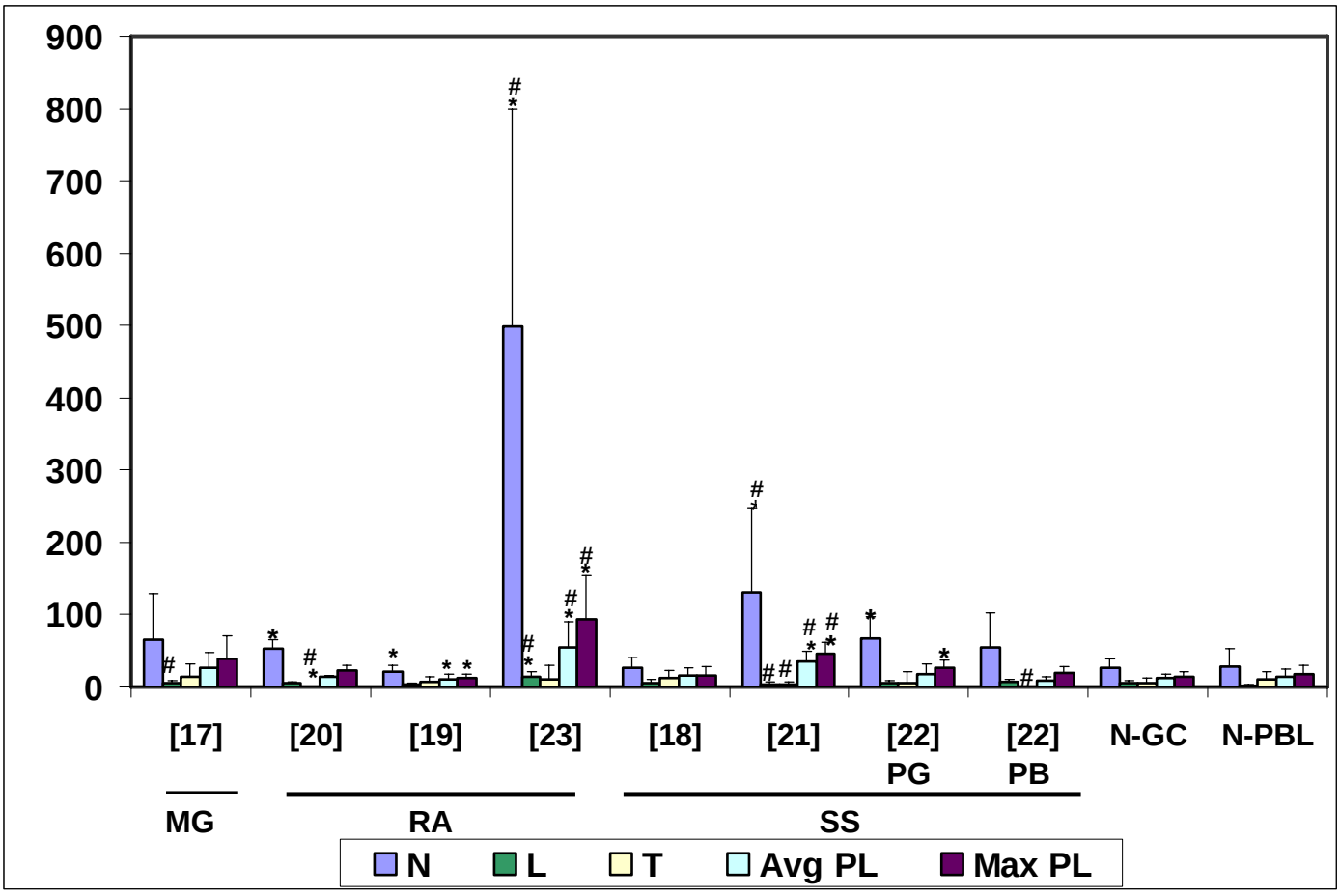
Figure 1. **(A)** A sample lineage tree. Nodes in the tree can be either the *root* node (always the node numbered zero in our definition, colored orange), *leaves* (sequences of cells that had no descendants at the time of observation – green nodes), or *internal nodes*. Internal nodes can be either *split nodes* – those with more than one child (yellow nodes); or *pass-through nodes* – those with exactly one child (white nodes). The trunk is the distance between the root to the first split node, which equals 1 in this case. **(B)** A comparison of the average number of nodes (N), the average number of leaves (L), the trunk length (T) and the maximum and average path length (MaxPL and AvgPL) in trees from the different experimental data groups. An * indicates $p<0.05$ in the comparison between the AI data group to the Normal GC data; an # indicates $p<0.05$ in the comparison between the AI data group and the Normal PBL data. Normal GC data (N-GC) is taken from [13,30], Normal PBL (N-PBL) data is taken from [15].

Figure 1.

(A)


Root
Trunk
Split Nodes
Pass-Through Nodes
0
1
2
3
4
5
6
7
8
9
10
11
12
13
Leaves

(B)


900
800
700
600
500
400
300
200
100
0
[17]
[20]
[19]
[23]
[18]
[21]
[22] PG
[22] PB
N-GC
N-PBL
MG
RA
SS
N
L
T
Avg PL
Max PL

Tree variable definition	**Abbreviation**	**Range**
Total number of nodes, including the root. Indicates the overall tree size.	N	$N \in [2, \infty)$
Total number of leaves, that is, the number of distinct sequences found, for which there were no "descendant" sequences.	L	$L \in [1, N-1]$
Number of internal nodes, that is, nodes that are not root or leaves.	IN	$IN \in [0, N-(L+1)]$
Number of pass-through nodes, that is, internal nodes that have only one child.	PTN	$PTN \in [0, IN]$
Length of tree trunk from root to the first split node, that is, the number of mutations shared by all leaves.	T	$T \in [0, N-1]$
Path length, where a path is defined from the root to a leaf, hence PL gives the number of mutations per leaf. Longer path lengths (in one group of trees relative to another) thus indicate that the cells are dividing more rapidly, and/or have a higher mutation rate (per division), and/or that the hypermutation process has been going on longer in these clones relative to those of the other group.	PL*	$PL \in [1, N-1]$
Distance from a leaf to the first (closest to root) split node: *DLFSN(leaf i) =PL(leaf i)-T*	DLFSN*	$DLFSN \in [1, MaxPL]$
Outgoing degree, representing the number of children per split node. Minimum and maximum OD are measured over split nodes, but if there are no splits (L=1) they both equal 1. AvgOD is measured over all nodes, including pass-through nodes. AvgOD2 represents the outgoing degree averaged only over all split nodes. ODs are measured of a tree's level of branching, or "bushiness", which is interpreted as indicating the rate of	OD*	$AvgOD \in [1,L]$ $MinOD \in [1,L]$ $MaxOD \in [1,L]$ $AvgOD2 \in [2,L]$

diversification relative to the strength of the selection forces acting on the tree, as selection tends to "prune" the tree (by killing cells with disadvantageous mutations) and hence reduce its bushiness.		
The root's outgoing degree, that is, the number of branches emerging from the root. *RootD=1⇔ T>0.*	RootD	*RootD ∈ [1,L]*
Distance between adjacent split nodes, that is, between two consecutive splits on the same path. It is an *inverse* measure of bushiness (or a direct measure of the relative strength of selection) over the whole tree, hence over the whole history of the clone.	DASN*	*DASN ∈ [1, N-(L+1)]*
Distance from a leaf to the last (closest to leaf) split node; an *inverse* measure of bushiness (or a direct measure of the relative strength of selection) over the *recent* history of the clone.	DLSN*	*DLSN ∈ [1,N-1]*
Distance from the root to any split node.	DRSN*	*DRSN ∈ [1, MaxPL]*

Table 1. Variables measured on each tree. Situations with no leaves (only a root node exists, that is, mutated sequences were not found, in which case N=1 and L=0) are not considered in our analysis. *For this variable we measure the maximum, minimum and average values (per tree).

Ref*	Disease	Tissue	# of Patients	# of Trees	Experimental Methods	<N>	Max L	Ig Chain	**Sequences found** [individual, rearranged productive, sequences / total sequences extracted]
[17]	MG	Thymus		13	Micro-dissection[i]	65.5	11	HC	18 B cell clones 92/216
[19]	RA	Synovial Tissue	2 (EK, PS)	4	Micro-dissection[ii]	19.7	5	HC, λLC & κLC	EK: VH:13 /40 κ:8/19 λ:6/19 PS: VH:9/25 κ:6/14 λ:2/2
[23]	RA	Synovial Tissue	2	20	Single cell suspensions	346.2	26	HC	263 IgG cells 276 IgA cells
[20]	RA	Synovial Tissue	2	4	Micro-manipulation	78.4	8	HC	63
[22]	SS	PB	1	5	Single cell suspensions	42.4	12	κLC	VκJκ:: 39/79 VλJλ: 54/81
		Parotid Gland		8		59.25	12	λLC & κLC	VκJκ:: 52/75 VλJλ: 29/38
[21]	SS	Labial salivary glands	2 (IL, UW)	10	Dissection	122.5	13	HC	IL-13/38 UW-13/14
		Lymph nodes		7		144	8	HC	IL-44/48
[18]	SS	Labial salivary glands	2 (BW, SG)	4	Micro-dissection	26	11	HC & LC	BW-22/27 SG-43/55

Table 2. Sources of data for analysis. *All trees were generated using our algorithm from sequence data, except those in [17-19] in which only trees and not sequences were given.

i - Serial sections 6-8 μm thick were extracted from 4 GC foci. Cells labeled for specifically for AchR as well as for CD20. Two rounds of PCR (35 and 40 cycles) performed, second cycle nested.

ii - 6 μm sections from consecutive sections extracted; In patient EK: 3 sections with 2 foci. In patient PS: 3 sections with 3 Foci. Sections stained for CD20, and KI-67 nuclear Ag. Two rounds of PCR (35 and 40 cycles) performed, second cycle semi-nested.

iii - Cells sorted for mononuclear cells. One round of PCR (35 cycles) performed.

iv - 7 µm sections with 12 infiltrate foci with 50-300 cells per focus were extracted. Sections stained for CD20. Two rounds of PCR performed.

v - Cells stained for CD19+. Two rounds of PCR (35 cycles each) performed, second round nested.

vi - 2 mm^3 section extracted from labial salivary gland, 60 mm^3 section extracted from lymph node. PCR (cDNA-35 cycles) performed. On salivary gland, in patient IL, 3 independent rounds and in patient UW, 2 independent rounds performed. On Lymph node 3 independent rounds performed.

vii - 8 µm sized sections. In patient BW three serial sections extracted from one cluster; In patient SG, 2 serial sections from 2 clusters. Cells stained for CD20 and anti CD3. Two rounds PCR (36 cycles), second round nested