# Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: chronic activation, normal selection

*Avital Steiman-Shimony*[a], *Hanna Edelman*[a], *Anat Hutzler*[a], *Michal Barak*[a], *Neta S. Zuckerman*[a], *Gitit Shahaf*[a], *Deborah Dunn-Walters*[b], *David I. Stott*[c], *Roshini S. Abraham*[d], *Ramit Mehr*[a,*]

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

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

[b]Division of Immunology, Infection and Inflammation, Glasgow Biomedical Research Centre, 120 University Place, University of Glasgow, Glasgow G12 8TA, Scotland, U.K.

[d]Division of Clinical Biochemistry and Immunology, Department of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, MN-55905 USA.

*****Correspondence: Prof. Ramit Mehr, Faculty of Life Sciences, Building 212, Box 61, Bar-Ilan University, Ramat-Gan 52900, ISRAEL. Phone: +972-3-531-7990. Fax: +972-3-535-1824. Email: mehrra@mail.biu.ac.il.

**ABSTRACT**

Autoimmune diseases show high diversity in the affected organs, clinical manifestations and disease dynamics. Yet they all share common features, such as the ectopic germinal centers found in many affected tissues. Lineage trees depict the diversification, via somatic hypermutation (SHM), of immunoglobulin variable-region (IGV) genes. We previously developed an algorithm for quantifying the graphical properties of IGV gene lineage trees, allowing evaluation of the dynamical interplay between SHM and antigen-driven selection in different lymphoid tissues, species, and disease situations. Here, we apply this method to ectopic GC B cell clones from patients with Myasthenia Gravis, Rheumatoid Arthritis, and Sjögren's Syndrome, using data scaling to minimize the effects of the large variability due to methodological differences between groups. Autoimmune trees were found to be significantly larger relative to normal controls. In contrast, comparison of the measurements for tree branching indicated that similar selection pressure operates on autoimmune and normal control clones.

**Keywords:** germinal centers, somatic hypermutation, Myasthenia Gravis (MG), Rheumatoid Arthritis (RA), Sjögren's Syndrome (SS).

## 1 INTRODUCTION

Germinal Centers (GCs) in primary lymphoid follicles are where rapid B cell proliferation, differentiation, somatic hypermutation (SHM) of immunoglobulin variable region (IGV) genes, and antigen-driven selection, lead to the preferential survival of those B cells with high affinity receptors to the antigen. Mutational lineage trees, depicting clonal relationships between related cells within a lineage, have frequently been drawn to illustrate IGV gene diversification in GC B cell clones derived from a few founder B cells. The qualitative features of IGV lineage trees have been used to interpret the dynamics of the GC response [1-2]. However, qualitative observations are limited to only the most obvious tree shape characteristics, and can only be used to compare a small number of trees at one time. Hence, we developed a rigorous computer-aided algorithm for the measurement of graphical shape properties of lineage trees, MTree© [3]. This method enables a more extensive investigation of the dynamics of the GC response. Previous studies have demonstrated the usefulness of this method, and resulted in new insights of various aspects of the GC reaction in normal situations [3-4], ageing [5-6], B cell malignancies [7,8], and chronic viral diseases (Margolin et al, submitted). In the present study, the GCs of individuals with autoimmune (AI) diseases are studied using lineage tree analysis.

The general cause and trigger of most AI diseases is unknown. Most AI disease studies focus on the role of T cells in the initiation of inflammation, whereas the production of pathogenic auto-antibodies by B cells is often overlooked [9]. Nevertheless, in many autoimmune diseases, ectopic GC-like areas develop within the afflicted tissue or organ, where AI B cells do undergo SHM and antigen mediated selection [10-14]. Hence, understanding the nature of ectopic GC B cell selection may yield new insights into the development of these diseases.

We created mutational lineage trees based on published IGV sequences from ectopic GC or affected tissue B cell clones in patients with Myasthenia Gravis (MG), Rheumatoid Arthritis (RA), Sjögren's Syndrome (SS), and Multiple Sclerosis (MS). These IGV lineage trees were analyzed using our MTree© program and compared to IGV trees from normal human samples. We observed a high variability between the different data sets [15] due to the data having been generated by different research groups at different times, using several methods to extract DNA sequences from samples of patients in which disease duration and severity probably varied as well. Moreover, the methods of tree generation from sequence alignments differed for some data sets. In the present study, we used a number of scaling techniques to alleviate the effects of experimental inconsistencies between research groups. Although the effects of the varying experimental methods were not completely neutralized, our analysis reveals that AI disease lineage trees are larger than their normal counterparts, likely as a result of the chronic nature of AI disease. Surprisingly, however, the selection process of the AI germinal center remains similar to that of the normal GC.

## 2 MATERIALS AND METHODS

### 2.1 Sources of sequences and database of IGV and tree data

A thorough literature and GenBank search yielded IGV sequences from patients with AI diseases. IGV sequences, lineage trees we created, and all tree measurements were stored in a database created specifically for lineage tree data storage and analysis, using a Microsoft SQL server (http://www.microsoft.com/sql/default.mspx), with a user-friendly Microsoft Access interface. Database tables store the administrative data of the experimental groups, the patients and the diseases, along with IGV sequences, the trees themselves and the resulting tree measurements. The tables are set up hierarchically,

using unique identification keys so that data in different tables can be connected and obtained through queries (Supplementary Figure 1), and allow for a convenient connection with statistical application programs.

### 2.2 Lineage tree generation and measurements

Some of the studies we found [10,12,14] contained lineage trees but no published sequences. In others, only sequences were published [16-21]; in these cases, germline genes were identified by alignment with published human IGV using DNAPLOT (http://vbase.mrc-cpe.cam.ac.uk/), and accordingly grouped and aligned using ClustalW (http://www.ebi.ac.uk/clustalw/). Trees were then generated using our program IgTree© (M. Barak et al, in preparation), which is specifically tailored to deal with IGV sequences. Lineage tree shape properties were quantified as described [15] by our MTree© program (Supplementary Figure 2, Supplementary Table 1).

### 2.3 Scaling

Tree properties were scaled by either the sequence alignment length (S), or by pick size – the total number of sequences, defined as either the "Tree" pick size (the number of sequences used per tree, not including the root sequence, but may have included identical sequences); or the "Lab group" pick size, i.e., the total number of sequences available in the study group that could be used to generate the trees.

### 2.4 Statistical analysis

GC simulation results (Shahaf *et al*. in preparation) show that lineage tree properties are not normally distributed. Hence we used the Mann-Whitney nonparametric test for

independent samples, and the Wilcoxon test for related samples. Multiple comparison correction (as we measured 25 different tree properties) was done by the FDR method [22], thus the minimal $\alpha$=0.05/25=0.002. FDR correction was performed separately for the unscaled and scaled data as they were compared independently of each other. Only differences that remained significant after FDR correction were considered as meaningful.

### 2.5 R:S Analysis

Replacement and silent mutations were enumerated and statistically analyzed with a computer program developed in our Lab (Zuckerman *et al*., unpublished) using the Lossos *et al.* multinomial correction [23] to the Chang & Casali method [24].

## 3 RESULTS

### 3.1 Large variability of AI data sets and the necessity for data scaling

Mutational lineage trees of B cell clones from patients with MG, RA, SS, and MS [6,16-21], and from normal human GCs [25-26], were created from published and unpublished IgV sequence data; we also used published lineage trees [12,10,14]. Each dataset contained IGV sequences from different IGV groups and patients (Table 1). Where lineage tree measurements of different groups within a dataset did not significantly differ, such groups were combined [15]. Sample trees are shown in Figure 1.

Tree measurements of AI data sets were compared to those of trees from normal human samples of both Peripheral blood lymphocytes (PBL [8]) and germinal centers (GC [6,26]), revealing large variability between the data of the different groups,

including between data sets of the same disease (Figure 2). Despite this variability, in all but two sets [10,14] the trees were larger than in the normal control data sets, whether we look at the number of leaves L, i.e. end cells in the sample (Figure 2A), the total number of mutations per clone, N (Figure 2B), or the maximum path length from root to leaf, PLmax (Figure 2C). This suggests a more vigorous diversification process in the ectopic AI GCs than in the normal controls.

The differences found between the L, N, and PL of the different AI datasets are most likely due to small methodological differences at all stages of data extraction (sampling, cell labeling, and DNA amplification), the source of tissue sample and the duration of the disease at time of sequence extraction. Moreover, tree generation methods may differ between data sets as some trees [10,12,14] were taken from published data where IGV sequences were not available. Two AI datasets [10,14] had smaller N and PL measures than those of the normal controls, probably also due to methodological differences in Ig sequence extraction and tree generation, as both these data groups come from the same lab.

In order to distinguish between methodological and actual differences, data were scaled by pick sizes or sequence alignment lengths. The sequence alignment length directly affects the number of nodes in a tree: the longer an alignment length, the greater the chances of finding mutations, hence longer alignments give large N and PL, amongst other outcomes. Thus N, PL, and partial paths on the tree (such as T, DASN or DLSN) may be scaled by sequence alignment length, to reduce the effects of sequence length differences, The pick size – the number of actual sequences which were sampled and are included in the tree – directly affects the number of leaves, which in turn affects the tree size N, and the degree of tree branching, as measured by

OD. Therefore, we scaled the N or OD measurements by either the "Tree pick size", or the "Lab group pick size", as defined above.

A comparison of the average sequence alignment lengths (Figure 3C) between data sets showed some variability, although slight, relative to that of AI datasets in the tree size measurements. On the other hand, the variability of the lab pick size and tree pick size (Figure 3 A and B) of the different AI data sets seemed to reflect the variability seen in the unscaled tree measurements.

### 3.2 Scaling confirms larger diversification in AI lineage trees

All measurements were scaled by the sequence alignment lengths and then the rates of mutation per division, as measured by the PL, or the related measure DLFSN, were compared (Figure 4). Although variability between the AI data sets persisted, most of their PL and DLFSN measurements remained significantly larger than those of the normal controls, supporting the hypothesis of a more vigorous diversification process in the AI germinal centers.

### 3.3 B cell selection in AI GCs is normal

While trees from AI diseases have longer paths, representing a longer diversification process, this may or may not indicate changes in B cell selection. The degree of tree branching, measured as the number of outgoing branches per node or outgoing degree (OD), indicates the extent of selection pressure, based on the assumption that selection will kill cells and thus reduce the number of leaves and hence of branches [4]. Unscaled OD measurements did not vary greatly between all different AI data sets, (not shown). Nonetheless, the experimental variability likely influenced OD measurements as well, hence they were scaled by tree pick size (number of sequences

sampled to create the tree), as the more sequences that are available to build a tree, the higher the likelihood of an increase in tree branching. A comparison of the scaled OD measurements (Figure 5) revealed that, in spite of the variability between AI data groups, the values remained of similar magnitude or smaller than those of normal controls. This suggested a trend of normal tree branching and correspondingly, a normal selection strength acting on the B cells in the ectopic germinal centers.

Similarly, the distances between adjacent split nodes (DASN), and the distances from a leaf to the last (or closest to the leaf) split node (DLSN), which are shorter in more branched trees and hence directly reflect selection strength, were scaled by sequence alignment length (Figure 6). The scaled DASN and DLSN, although slightly less variable than their unscaled values, were still mostly as large as those of control tissues, confirming the hypothesis of normal selection pressure.

As an additional verification, we performed R:S mutation analysis [23-29] on our data, showing that the AI clones underwent positive antigen-mediated selection, at least as strong as that in normal clones (Supplementary Table 3). R:S analysis is known to give many false positive indications of selection [30-31]. Nonetheless, it may be used to validate the existence of selection in cases where it has already been suggested by our lineage analysis, though it cannot be used to refute our findings in cases where we found no selection.

### 3.4 MS data from two time points

Sequences obtained from two MS patients [20] were taken at two points in time each: Patient 1 was sampled at disease onset (time 0) and 1 year later (time 1); Patient 2 was sampled 9 years (time 0) and 13 years (time 1) after initial Diagnosis. Sequences from the two time points were compared in [20] for each patient using oligoclonal

patterns, heavy chain CDR3 lengths – which are a measure of repertoire diversity in normal populations – and frequencies of VH families used. The study concluded that a more active diversification pattern occurred in the initial stages of the disease followed by a persistence of "memory" clones that were selected for their antigenic specificities.

In order to better understand the diversification and selection pressures which occur over time in MS, we created IGV lineage trees from published sequences [20]. The tree measurements of those clones from the same patient that appeared in both time points were compared to each other, and a paired Wilcoxon test was performed. Because only four clones from both time points were available for each patient, tests lacked statistical significance. Nevertheless interesting trends were observed. although the number of accumulated mutations per cell, as measured by PL, increased from the first time point to the second in both patients, as would be expected if the same clones continued to develop with time, the degree of tree branching, represented by the maximum OD value (Figure 7), remained similar. Thus, in contrast to the conclusion of [20], our analysis suggests similar diversification patterns at both time points for both patients. Moreover, an increase in selection pressure in patient one and a decrease in patient two was shown by the DASN and DLSN measurements (Figure 7). The consequences which led to the observed changes are unknown, and more clones and time points would be necessary to make clear conclusions as to the nature of the B cells in the ectopic germinal centers of MS patients.

## 4 DISCUSSION

The present study aimed to investigate the ontogeny and progress of AI diseases, using lineage tree analysis of Ig sequences. To reduce the variability caused by

methodological differences between experimental groups, data were scaled by pick size or sequence alignment length. The results depicted the larger sizes of AI trees relative to the normal controls, indicating more mutations accumulated and a more vigorous diversification process, as expected due to the chronic nature of AI diseases. Even more notable was the unexpected discovery that selection acting on AI disease cells was no different from that in controls, despite the irregular locations of ectopic GCs. This conclusion is significant because, while the oligoclonal nature of the sequences in the germinal center was apparent from the experimental studies from which data were extracted, the extent of selection, if any, was unclear. In this context it is noteworthy that follicular dendritic cells have been identified in ectopic GCs in RA [32], SS [10], and MG [12] and their presence supports the evidence for antigen-driven selection.

The study by Colombo et al., [20], where Ig sequences were extracted from MS patients at two time points in the course of their disease, concluded that a more active diversification pattern occurs in the initial stages of the disease. The analysis presented here does not support this conclusion, as diversification shown by the PL measures continues at a similar pace. Moreover, the selection pressures in both cases seem to operate similarly on the clones.

Lineage tree analysis thus has the potential of yielding new findings concerning ectopic GCs in MS and other AI diseases. In order to successfully do so, in future studies larger Ig sequence sampling, the sampling of more than two time points, and the study of more patients must be undertaken. Furthermore, studies in which sequences are extracted from a larger number of patients with the same disease undergoing different treatments, or varying in symptoms or response to treatment, at different time points, may lead to an understanding of the progression of the diseases

and the effects of their treatments. In most Ig sequence data available to date, only the V segment of the Ig gene is sequenced, and hence in this study only the V segments of the Ig sequences were aligned for tree generation. As a result, trees are smaller than they would have been had the whole IGV segment been used.

## ACKNOWLEDGEMENTS

The authors are indebted to Osnat Steiman for editing the manuscript. The work was part of Avital Steiman-Shimony's studies towards an MSc degree in Bar-Ilan University, and was supported in parts by the following grants: the Israel Science Foundation grants number 759/01-1 and 546, an Israel Cancer Research Fund project grant, a Systems Biology prize grant from Teva Pharmaceuticals, a Human Frontiers Science Program Young Investigator Grant, and a Swedish Foundation for Strategic Research grant funding the Strategic Research Center for studies on Integrative Recognition in the Immune System (IRIS), Karolinska Institute, Stockholm, Sweden (to RM); a BBSRC Science of Ageing initiative grant (to DDW); grants from the EC Marie Curie Research Fellowship (the SS work), and The Wellcome Trust (the MG work), to DIS; and a Hematological Malignancies Research Fund, Mayo Clinic grant (to RSA).

## REFERENCES


[1] J. Jacob, G. Kelsoe, 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 176 (1992) 679-687.

[2] J. Jacob, J. Przylepa, C. Miller, G. Kelsoe, 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. 178 (1993) 1293-1307.

[3] D.K. Dunn-Walters, A. Belelovsky, H. Edelman, M. Banerjee, R. Mehr R, The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees, Dev. Immunol. 9 (2202) 233-243.

[4] D.K. Dunn-Walters, H. Edelman, R. Mehr, Immune system learning and memory quantified by graphical analysis of B-lymphocyte phylogenetic trees, BioSystems 76 (2004) 141-155.

[5] M. Banerjee, R. Mehr, A. Belelovsky, J. Spencer, D.K. Dunn-Walters, Age and tissue specific differences in human germinal center B cell selection revealed by analysis of IGVH gene hypermutation and lineage trees, Eur. J. Immunol. 32 (2002) 1947-1957.

[6] D.K. Dunn-Walters, M. Banerjee, R. Mehr, Age effects on antibody affinity maturation, Biochem, Soc. Trans. 31 (2003) 447-448.

[7] M.K. Manske, N.S. Zuckerman, M.M. Timm, S. Maiden, H. Edelman, G. Shahaf, M. Barak, A. Dispenzieri, M.A. Gertz, R. Mehr, R.S. Abraham, Clonal CD19+ B cells skew the Ig VL gene repertoire in the CD138-negative compartment of bone marrow in light chain Amyloidosis, Clinical Immunology 120 (2006) 106-120.

[8] R.S. Abraham, M.K. Manske, N.S. Zuckerman, A. Sohni, H. Edelman, G. Shahaf, M.M. Timm, A. Dispenzieri, M.A. Gertz, R. Mehr, Novel Analysis of Clonal

Diversification in Blood B Cell and Bone Marrow Plasma Cell Clones in Immunoglobulin Light Chain Amyloidosis, J. Clin. Immunol (2007) in press.

[9] P. Youinou, C. Jamin, J.O. Pers, C. Berthou, A. Saraux, Y. Renaudineau, B Lymphocytes Are Required for Development and Treatment of Autoimmune Diseases, Ann. N.Y. Acad. Sci. 1050 (2005) 19-33.

[10] D.I. Stott, F. Hiepe, M. Hummel, G. Steinhauser, C. Berek, 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. 102 (1998) 938-946.

[11] J. William, C. Euler, S. Christensen, M.J. Shlomchik, Evolution of Autoantibody Responses via Somatic Hypermutation Outside of Germinal Centers, Science, 297 (2002) 2066-2070.

[12] G.P. Sims, H. Shiono, N. Willcox, D.I. Stott, Somatic Hypermutation and Selection of B Cells in Thymic Germinal Centers Responding to Acetylcholine Receptor in Myasthenia Gravis, J. Immunol. 16 (2001) 1935-1944.

[13] B. Serafini, B. Rosicarelli, R. Magliozzi, E. Stigliano, F. Aloisi, Detection of Ectopic B-cell Follicles with Germinal Centers in the Meninges of Patients with Secondary Progressive Multiple Sclerosis, Brain Pathol. 14 (2004) 164-174.

[14] H.J. Kim, V. Krenn, G. Steinhauser, C. Berek, Plasma Cell Development in Synovial Germinal Centers in Patients with Rheumatoid and Reactive Arthritis, J. Immunol. 162 (1999) 3053-3062.

[15] A. Steiman-Shimony, H. Edelman, M. Barak, G. Shahaf, D.K. Dunn-Walters, D.I. Stott, R.S. Abraham, R. Mehr, Immunoglobulin variable-region gene mutational lineage tree analysis: Application to autoimmune diseases, Autoimmun Rev. 5 (2006) 242-251.

[16] M. Colombo, M. Dono, P. Gazzola, S. Roncella, A. Valetto, N. Chiorazzi, G.L. Mancardi, M. Ferrarini, Accumulation of clonally related B lymphocytes in the cerebrospinal fluid of multiple sclerosis patients, J. Immunol. 164 (2000) 2782-2789.

[17] Y. Miura, C.C. Chu, D.M. Dines, S.E. Asnis, R.A. Furie, N. Chiorazzi, Diversification of the Ig Variable Region Gene Repertoire of Synovial B Lymphocytes by Nucleotide Insertion and Deletion, Mol. Med. 9 (2003) 166-174.

[18] A. Gause, K. Gundlach, M. Zdichavsky, G. Jacobs, B. Koch, T. Hopf, M. Pfreundschuh, The B lymphocyte in rheumatoid arthritis: analysis of rearranged V kappa genes from B cells infiltrating the synovial membrane, Eur. J. Immunol. 25 (1995) 2775-2782. Data published in Genbank (GenBank accession nos X85162-X85167, Z75254, Z75333-Z75343, Z75442-Z75465).

[19] S. Gellrich, S. Rutz, A. Borkowski, S. Golembowski, E. Gromnica-Ihle, W. Sterry, S. Jahn, 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. 42 (1999) 240-247.

[20] M. Colombo, M. Dono, P. Gazzola, N. Chiorazzi, G. Mancardi, M. Ferrarini, Maintenance of B lymphocyte realted clones in the cerebrospinal fluid of multiple sclerosis patients, Eur. J. Immunol. 33 (2003) 3433-3438.

[21] A.M. Jacobi, A. Hansen, O. Kaufmann, A. Pruss, G.R. Burmester, P.E. Lipsky, T. Dorner, Analysis of immunoglobulin light chain rearrangements in the salivary gland and blood of a patient with Sjögren's syndrome, Arthritis Res. 4 (2002) R4.

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

[23] I.S. Lossos, R. Tibshirani, B. Narasimhan, R. Levy, The inference of Ag selection on Ig genes, J. Immunol. 165 (2000) 5122 – 5126.

[24] B. Chang, P. Casali, The CDR1 sequences of a major proportion of human germline Ig VH genes are inherently susceptible to amino acid replacement, Immunol. Today. 15 (1994) 367-373.

[25] R.S. Abraham, K.V. Ballman, A. Dispenzieri, D.E. Grill, M.K. Manske, T.L. Price-Troska, N.G. Paz, M.A. Gertz, R. Fonseca, Functional gene expression analysis of clonal plasma cells identifies a unique molecular profile for light chain amyloidosis, Blood. 105 (2005) 794-803

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

[27] T. Dorner, S.J. Foster, N.L. Farner, P.E. Lipsky, Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands, Eur. J. Immunol. 28 (1998) 3384-3396.

[28] D.I. Stott, C. Berek, An antigen-driven B-cell response within the salivary glands of patients with Sjögren's syndrome, Annals of the Marie Curie Fellowship Association. 2 (2002) 108 – 116.

[29] S. Nzula, J.J. Going, D.I. Stott, Antigen-driven Clonal Proliferation, Somatic Hypermutation, and Selection of B Lymphocytes Infiltrating Human Ductal Breast Carcinomas, Cancer Res. 63 (2003) 3275-3280.

[30] D.K. Dunn-Walters, J. Spencer, Strong intrinsic biases towards mutation and conservation of bases in human IGVH genes during somatic hypermutation prevent statistical analysis of antigen selection, Immunology. 95 (1998) 339-345.

[31] B. Bose, S. Sinha, Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection, Immunology. 116 (2005) 172–183.

[32] A.E. Schroder, A. Greiner, C. Seyfert, C. Berek, Differentiation of B cells in the nonlymphoid tissue of the synovial membrane of patients with rheumatoid arthritis, Proc. Natl. Acad. Sci. 93 (1996) 221-225.

## FIGURE LEGENDS

**Figure 1.** Sample tree figures created from B cell clones extracted from: A. Normal Peyer's patch [5]; B. Light chain B cell clones from the parotid gland of a SS patient [21]; C. Heavy chain B cell clones from a lymph node of a SS patient [19]; D. Heavy chain B cell clones from the CSF of an MS patient [16]. Double circles represent the germline (unmutated) root sequence; the dark circles are experimentally generated sequences, most of which are leaves; the dashed circles represent deduced intermediate sequences. The numbers along the vertical lines represent the number of mutations unique to a sequence compared to the immediately preceding sequence.

**Figure 2 .** Comparison of AI disease data to normal controls, unscaled: **A)** Number of leaves, L; **B)** Number of nodes, N; **C)** maximum path length, PLmax. **Data sources** are as in Table 1. Significant differences between AI datasets to normal GC control are labeled by '*'; Significant differences between AI datasets to normal PBL controls are labeled by '#'. P Values are given in Supplementary Table 2(A). Tree properties are as described in the text and in Supplementary Table 1.

**Figure 3.** Comparison of all data sets for Lab group pick size (A), and Tree pick size (B) – both given in numbers of sequences; and sequence alignment length (C), given in number of nucleotides (except for groups 1 and 5, for which we do not have the sequence alignment lengths). Data sources are as in Figure 2. Tree properties are as described in the text and in Supplementary Table 1.

**Figure 4.** Comparison of maximum PL (white) and DLFSN (striped) measures of AI data sets to normal controls, scaled by the sequence alignment length (minimum and

average measures reveal similar trends). Data sources and significance markers are as in Figure 2; P Values are given in Supplementary Table 2(B). Tree properties are as described in the text and in Supplementary Table 1.

**Figure** 5 Comparison of maximum L (gray) and maximum OD (striped) measures of AI data sets compared to normal controls, scaled by the tree pick size (minimum and average OD measures reveal similar trends). Data sources and significance markers are as in Figure 2; P Values are given in Supplementary Table 2(C). Tree properties are as described in the text and in Supplementary Table 1.

**Figure 6.** Comparison of maximum DASN (white) and DLSN (striped) measures, unscaled (top) and scaled by sequence alignment length (bottom), of the AI data sets to the normal controls. Data sources and significance markers are as in Figure 2; P Values are given in Supplementary Table 2(D,E). Tree properties are as described in the text and in Supplementary Table 1.

**Figure 7.** Comparison between time point 1 (white) and 2 (striped) of tree measurements T, L, as well as maximum PL, OD, DASN and DLSN values (similar trends were found for average values of these measurements), from patient 1 (top) and patient 2 (bottom). Tree properties are as described in the text and in Supplementary Table 1.

**Supplementary Figure 1**: The Hierarchical structure of the IGV lineage tree database.

**Supplementary Fig 2** A sample Lineage tree (left) and the same lineage tree in text format as tree list (right). A lineage tree is defined, graphically, as a rooted tree where the nodes correspond to B cell receptor gene sequences. 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 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 complete list of variables measured is given in **Supplementary Table 1**, including remarks on the meaning and usefulness of each variable.

**FIGURES**

**Figure 1:**

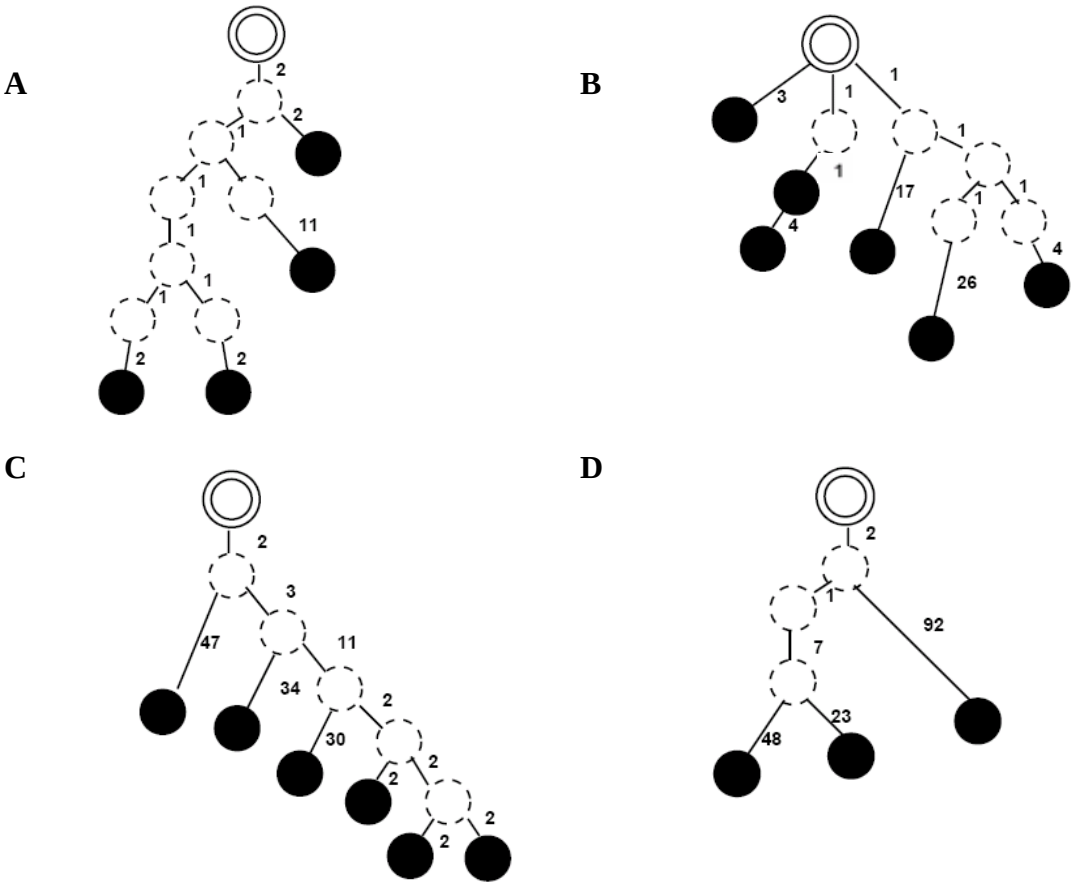



**Figure 2:**

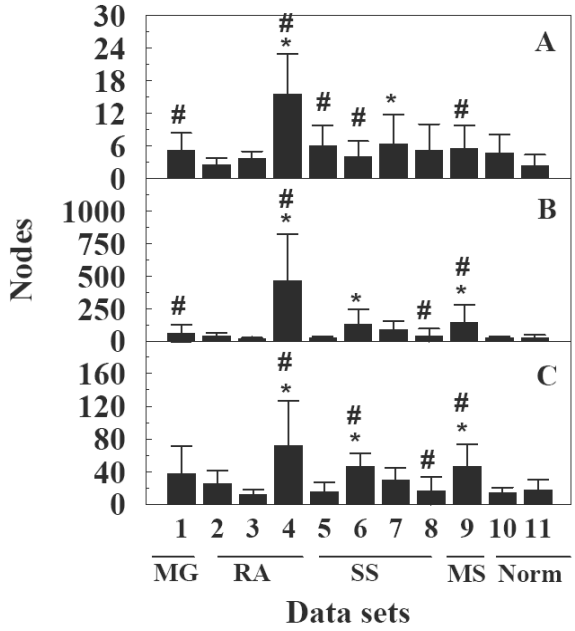


**Figure 3:**

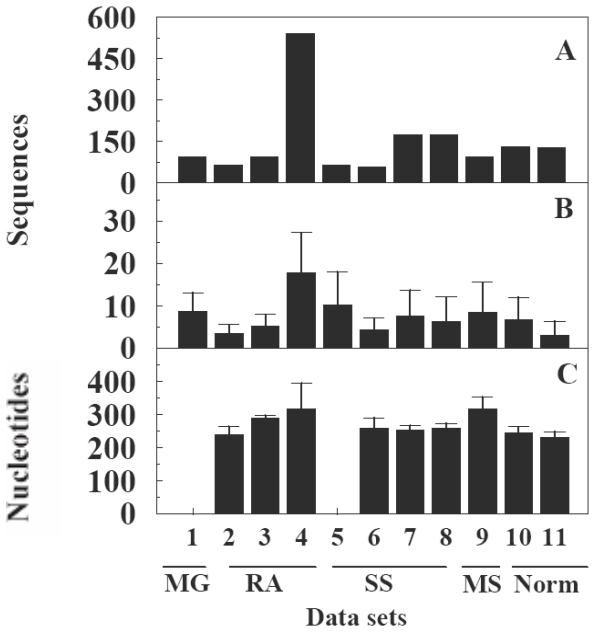



**Figure 4:**

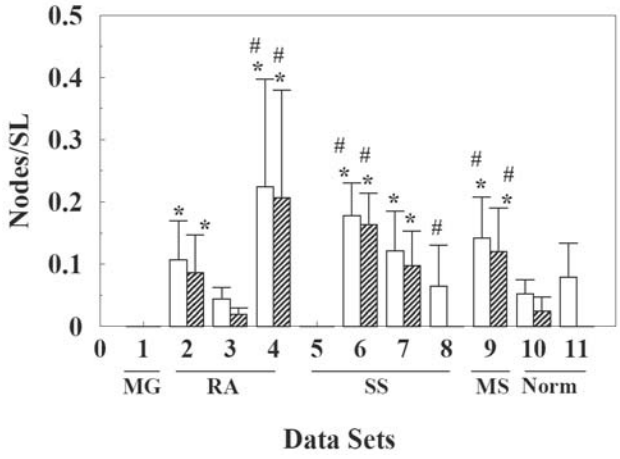



**Figure 5:**

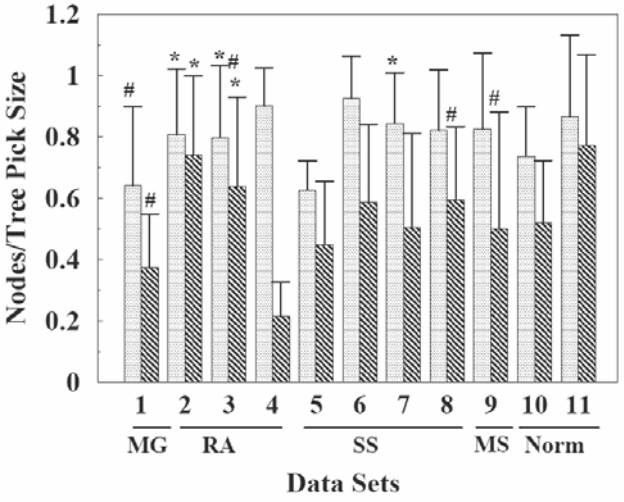


**Figure 6:**

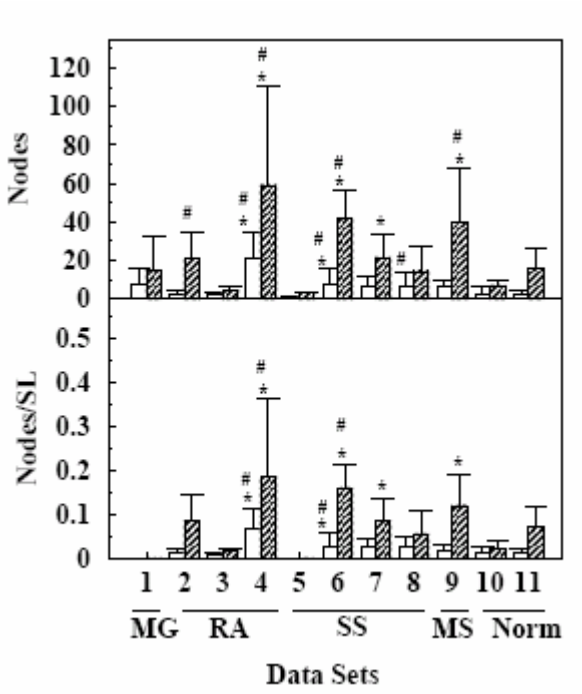
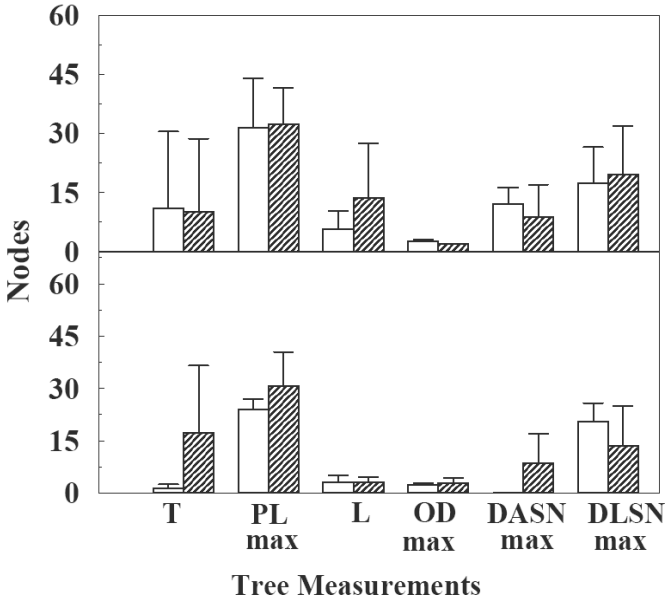
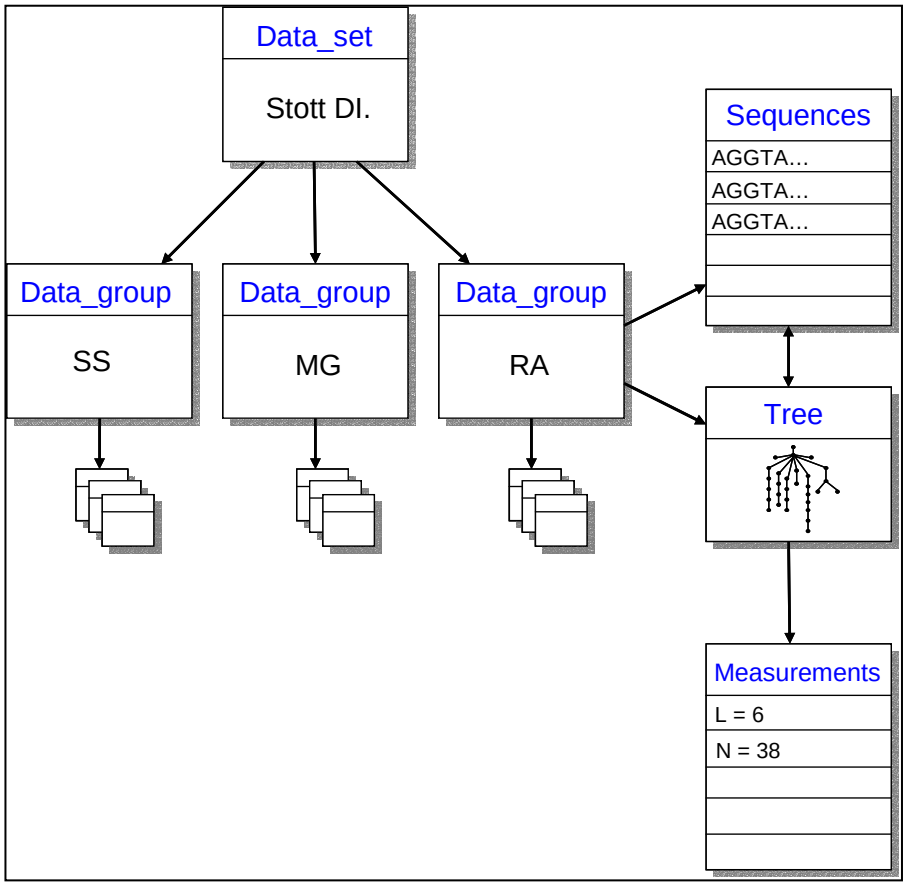


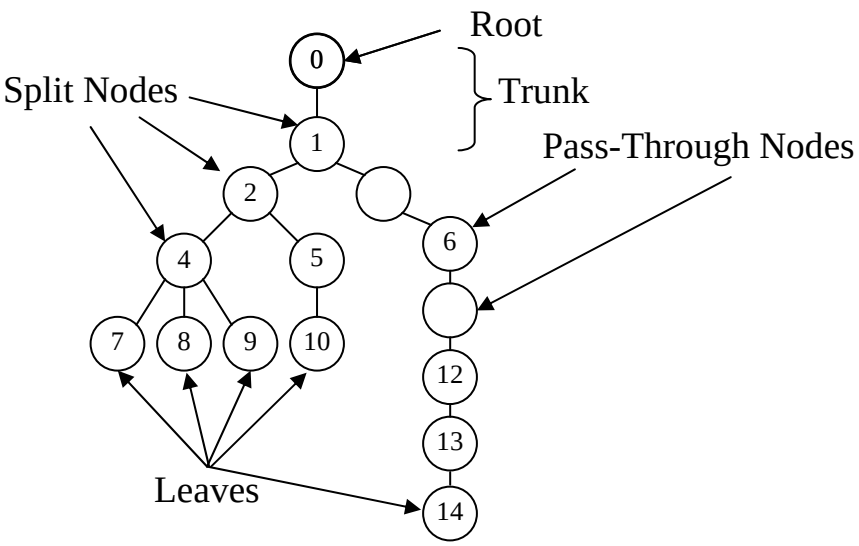

**Figure 7:**



**Supplementary Figure 1:**




**Supplementary Figure 2:**




This tree as a list:

| | | | |
|---|---|---|---|
| 0 | 1 | | |
| 1 | 2 | 3 | |
| 2 | 4 | 5 | |
| 3 | 6 | | |
| 4 | 7 | 8 | 9 |
| 5 | 10 | | |
| 6 | 11 | | |
| 11 | 12 | | |
| 12 | 13 | | |
| 13 | 14 | | |

**Table 1**: **Data sources**

| Ref[a] | Disease / Tissue | # of Patients | # of Trees | Methods | Ig Chain |
|---|---|---|---|---|---|
| 1 | MG / Thymus | 1 | 13 | MD | HC |
| 2 | RA / ST | 2 | 12 | MD | HC |
| 3 | RA / ST | 2 | 4 | MD | HC, λLC, κLC |
| 4 | RA / ST | 2 | 14 | SCS | HC |
| 5 | SS / LSG | 2 | 4 | MD | HC, LC |
| 6 | SS / LSG | 2 | 10 | MD | HC |
| | SS / LN | | 7 | | |
| 7, 8 | SS / PBL | 1 | 12 | SCS | κLC |
| | SS / PG | | 9 | | λLC, κLC |
| 9 | MS / CSF | 2 | 7 | SCS | HC |
| | MS / PBL | | 3 | | |
| 10 | Normal / LN | 2 | 4 | MD | HC |
| | Normal / PP | 5 | 7 | MD | HC |
| | Normal / Spleen | | 12 | | |
| 11 | Normal / PBL | 2 | 42 | SCS | LC |

1-Sims et al. 2001 [12], 2-Gause et al. 1995 [18], 3-Kim et al. 1999 [14], 4-Miura et al., 2003 [17], 5-Stott et al., 1998 [10], 6-Gellrich et al., 1999 [19], 7-PG data; Jacobi et al., 2002 [21], 8-PBL data; Jacobi et al., 2002[21], 9-Colombo et al., 2000 [16], 10-Kuppers et al. [21], 1993, and Banerjee et al., 2002 [5], 11- Abraham et al. 2005[22]. **Tissues :** ST : Synovial tissue ; PG : Parotid Gland; LSG; Labial salivary glands; LN: Lymph nodes; PP: Peyer's Patch. **Experimental methods:** MD: Micro-dissection; SCS: Single cell suspensions. [a]All trees were generated using our algorithm from sequence data, except those in Sims et al., 2001 [12], Stott et al., 1998 [10], and Kim et al., 1999 [14], in which only trees and not sequences were given.

**Supplementary Table 1**: Tree measurements and their definitions

| **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.[b] | PL[a] | $PL \in [1, N-1]$ |
| Distance from a leaf to the first (closest to root) split node: *DLFSN(leaf i) =PL(leaf i)-T* | DLFSN[a] | $DLFSN \in [1, MaxPL]$ |
| Outgoing degree, representing the number of children per split node.[c] | OD[a] | $AvgOD, MinOD, and\ MaxOD \in [1,L]$<br>$AvgOD2 \in [2,L]$ |
| The root's outgoing degree, that is, the number of branches emerging from the root. $RootD=1 \Leftrightarrow T>0$. | RootD | $RootD \in [1,L]$ |
| Distance between adjacent split nodes, that is, between | DASN[a] | $DASN \in [1, N-(L+1)]$ |

| | | |
|---|---|---|
| two consecutive splits on the same path.[d] | | |
| Distance from a leaf to the last (closest to leaf) split node.[e] | DLSN[a] | *DLSN ∈ [1,N-1]* |
| Distance from the root to any split node. | DRSN[a] | *DRSN ∈ [1, MaxPL]* |

a. Minimum, Maximum and Average values measured

b. 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.

c. 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 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.

d. 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.

e. It is an *inverse* measure of bushiness (or a direct measure of the relative strength of selection) over the *recent* history of the clone.

**Supplementary Table 3: R:S analysis[a]**

| Data Set | # of Sequences Checked | Pval Sig in FR | Pval Sig in CDR | R/S <2.9 in FR | R/S>=2.9 in CDR |
|---|---|---|---|---|---|
| 2 | 40 | 12/40<br>3.00E-01 | 12/40<br>3.00E-01 | 29/40<br>7.25E-01 | 20/40<br>5.00E-01 |
| 4 | 191 | 143/191<br>7.49E-01 | 94/191<br>4.92E-01 | 179/191<br>9.37E-01 | 90/191<br>4.71E-01 |
| 6 | 74 | 52/74<br>7.03E-01 | 42/74<br>5.68E-01 | 69/74<br>9.32E-01 | 39/74<br>5.27E-01 |
| 7 | 72 | 34/72<br>4.72E-01 | 29/72<br>4.03E-01 | 64/72<br>8.89E-01 | 30/72<br>4.17E-01 |
| 8 | 63 | 20/63<br>3.17E-01 | 15/63<br>2.38E-01 | 60/63<br>9.52E-01 | 18/63<br>2.86E-01 |
| 9 | 101 | 67/101<br>6.63E-01 | 55/101<br>5.45E-01 | 76/101<br>7.52E-01 | 47/101<br>4.65E-01 |
| DDW | 129 | 11/129<br>8.53E-02 | 18/129<br>1.40E-01 | 93/129<br>7.21E-01 | 52/129<br>4.03E-01 |

[a]A comparison of the results between the different AI datasets to the normal controls supported the results already shown by the OD tree measurements from our study. The analysis showed (third column) that the number of IGV sequences in each of the corresponding AI datasets had statistically significant differences ($p < 0.05$) from the random diversification frequencies of R and S mutations in the FR. All AI datasets were found to have higher fractions than those in the normal datasets analyzed,

indicating that selection occurred in at least some of the AI clones. Similarly, the AI dataset clones had larger fractions of statistically significant differences ($p < 0.05$) from the random diversification frequencies of R and S mutations in the CDR than the normal control datasets' clones (fourth column). Again, this indicated that the AI clones analyzed had undergone selection.

The R:S mutation ratio expected under a random process based on the inherent mutability of all codons and their frequencies is 2.9; hence an R:S ratio less than 2.9 in the FR and at least 2.9 in the CDR is believed to represent positive antigen mediated selection of the B cell clones analyzed[b]. The fraction of sequences to have an R:S ratio less than 2.9 (or in the case where S=0, R is less than 2.9) in the FR is shown in the fifth column. The results indicated that all of the AI datasets had the same or higher fractions of clones with an R:S ratio less than 2.9 in the FR. The sixth column which contains the fractions of clones with an R:S ration more, or equal to 2.9 (or in the case where S=0, R is at least 2.9) in the CDR, resulted in similar indications. All AI datasets had higher fractions of clones with an R:S ratio equal or above the random ration of 2.9. Hence, it was apparent that the AI clones underwent positive antigen-mediated selection, at least as strong as that

in normal clones, which supported our previous findings using tree measurement data.

[b]Jukes T.H., and King J.L. 1979. Evolutionary nucleotide replacements in DNA. *Nature 281:* 605-606.

## Supplementary Table 2: P-values from comparisons of tree properties

| | AI Datasets | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Control sets | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 | 10 | 11 |
| A | L | 3.97E-01 | 7.96E-05 | 2.12E-02 | 7.33E-02 | 8.69E-01 | 3.73E-02 | 4.74E-06 | 1.60E-07 | 4.89E-01 | 7.76E-03 | 6.85E-01 | 3.04E-04 | 4.83E-01 | | 2.07E-03 | | 5.50E-01 | 9.58E-04 |
| | N | 2.81E-01 | 1.20E-01 | 7.91E-02 | 3.31E-02 | 6.69E-01 | 8.45E-01 | 8.66E-10 | 7.83E-08 | 1.00E+00 | 6.82E-01 | 8.41E-09 | 2.63E-07 | 6.00E-04 | | 5.88E-01 | | 6.48E-07 | 2.90E-05 |
| | PLmax | 8.69E-02 | 8.64E-02 | 2.12E-02 | 9.78E-02 | 7.18E-01 | 4.35E-01 | 8.66E-10 | 6.84E-07 | 7.18E-01 | 8.61E-01 | 1.58E-10 | 2.86E-07 | 2.57E-03 | | 3.70E-01 | | 2.10E-06 | 2.76E-04 |
| B | PLmin | NA | NA | 6.02E-01 | 5.28E-01 | 1.00E+00 | 5.68E-01 | 6.81E-02 | 3.88E-01 | NA | NA | 1.54E-06 | 1.99E-03 | 2.12E-01 | 9.56E-02 | 9.56E-04 | 3.35E-04 | 6.86E-06 | 1.41E-02 |
| | PLmax | NA | NA | 4.99E-03 | 1.05E-01 | 6.52E-01 | 1.70E-01 | 9.70E-09 | 3.68E-06 | NA | NA | 4.41E-10 | 8.96E-07 | 1.54E-03 | 5.19E-02 | 8.84E-01 | 1.99E-01 | 2.12E-05 | 5.55E-03 |
| | PLav | NA | NA | 1.50E-02 | 3.28E-01 | 7.12E-01 | 2.10E-01 | 3.88E-08 | 2.02E-04 | NA | NA | 8.82E-10 | 3.87E-06 | 8.89E-03 | 5.05E-01 | 1.04E-01 | 8.07E-03 | 1.48E-05 | 5.17E-03 |
| | DLFSNmin | NA | NA | 5.18E-01 | 2.11E-01 | 9.82E-02 | 3.72E-01 | 1.04E-03 | 1.48E-03 | NA | NA | 7.03E-07 | 1.28E-05 | 7.43E-01 | 5.09E-01 | 3.87E-01 | 9.85E-01 | 2.90E-06 | 1.23E-04 |
| | DLFSNmax | NA | NA | 5.78E-03 | 3.30E-01 | 6.59E-02 | 5.87E-02 | 9.70E-09 | 1.72E-05 | NA | NA | 4.41E-10 | 5.22E-06 | 6.25E-03 | 1.33E-01 | 1.73E-01 | 9.27E-01 | 4.10E-05 | 2.31E-02 |
| | DLFSNav | NA | NA | 1.16E-02 | 1.35E-01 | 8.07E-02 | 5.87E-02 | 9.70E-09 | 1.48E-05 | NA | NA | 4.41E-10 | 7.87E-07 | 1.45E-01 | 5.20E-03 | 6.11E-01 | 4.96E-01 | 2.12E-05 | 2.56E-03 |
| C | L (Leaves) | 2.93E-01 | 4.72E-03 | 2.67E-01 | 4.18E-01 | 8.37E-01 | 6.99E-01 | 4.86E-03 | 5.87E-01 | 1.95E-01 | 9.84E-02 | 1.09E-03 | 3.60E-01 | 5.95E-02 | 4.17E-01 | 1.46E-01 | 3.55E-01 | 7.17E-02 | 5.93E-01 |
| | ODmin | 2.26E-01 | 1.18E-04 | 1.32E-02 | 5.35E-01 | 9.67E-01 | 2.40E-01 | 2.91E-04 | 2.59E-07 | 7.74E-01 | 4.23E-02 | 1.26E-01 | 3.83E-02 | 8.60E-01 | 2.11E-02 | 6.02E-01 | 1.28E-02 | 9.81E-01 | 2.43E-02 |
| | ODmax | 5.64E-02 | 5.66E-05 | 1.50E-02 | 5.96E-01 | 3.42E-01 | 3.26E-01 | 1.52E-05 | 2.16E-07 | 7.12E-01 | 4.23E-02 | 3.69E-01 | 2.19E-02 | 6.68E-01 | 2.77E-02 | 3.46E-01 | 5.89E-02 | 5.88E-01 | 1.63E-02 |
| | ODavg | 1.46E-01 | 1.88E-04 | 3.46E-02 | 6.67E-02 | 7.12E-01 | 1.47E-01 | 1.98E-05 | 6.04E-07 | 4.84E-01 | 1.45E-02 | 4.05E-01 | 8.59E-04 | 4.63E-01 | 5.72E-03 | 4.91E-01 | 2.37E-03 | 7.24E-01 | 2.05E-03 |
| | ODavg2 | 1.35E-01 | 6.74E-03 | 1.32E-02 | 9.84E-01 | 6.52E-01 | 5.92E-01 | 5.29E-05 | 2.26E-05 | 9.02E-01 | 2.62E-01 | 1.43E-01 | 2.68E-01 | 6.31E-01 | 1.25E-01 | 6.02E-01 | 1.69E-01 | 7.96E-01 | 1.10E-01 |
| D | DASNmax | 1.23E-01 | 3.36E-01 | 6.48E-01 | 5.56E-01 | 7.05E-01 | 8.83E-01 | 2.93E-06 | 5.31E-04 | 6.69E-01 | 1.18E-01 | 1.02E-02 | 2.72E-02 | 7.06E-02 | | | 3.67E-01 | 2.15E-02 | 2.98E-02 |
| | DASNav | 1.00E-01 | 2.61E-01 | 7.05E-01 | 1.25E-03 | 4.42E-01 | 9.42E-01 | 1.21E-02 | 3.13E-02 | 6.69E-01 | 1.18E-01 | 8.20E-03 | 4.90E-02 | 3.56E-01 | | | 9.69E-01 | 6.90E-02 | 4.70E-01 |
| | DLSNmax | 6.03E-01 | 2.60E-01 | 5.68E-04 | 7.33E-02 | 2.43E-01 | 1.29E-02 | 8.66E-10 | 2.40E-05 | 2.75E-02 | 8.17E-03 | 2.25E-11 | 1.28E-05 | 1.54E-03 | | | 3.40E-01 | 1.45E-06 | 7.16E-03 |
| | DLSNav | 8.45E-01 | 2.02E-02 | 2.97E-04 | 2.51E-01 | 4.09E-01 | 1.63E-02 | 8.66E-10 | 6.91E-05 | 1.38E-02 | 4.15E-03 | 2.25E-11 | 4.93E-07 | 4.90E-04 | | | 1.32E-01 | 5.81E-06 | 2.23E-03 |
| E | DASNmax | NA | NA | 8.87E-01 | 9.46E-01 | 9.62E-01 | 8.00E-01 | 4.18E-05 | 5.32E-04 | NA | NA | 3.10E-02 | 2.74E-02 | 1.37E-01 | 2.10E-01 | 2.16E-01 | 4.46E-01 | 1.41E-01 | 5.19E-02 |
| | DASNav | NA | NA | 8.87E-01 | 7.36E-01 | 7.40E-01 | 1.00E+00 | 8.90E-02 | 1.91E-01 | NA | NA | 7.70E-03 | 2.75E-02 | 6.73E-01 | 8.31E-01 | 5.67E-01 | 9.13E-01 | 2.97E-01 | 4.51E-01 |
| | DLSNmax | NA | NA | 1.37E-03 | 3.69E-01 | 6.59E-02 | 1.12E-02 | 9.70E-09 | 1.61E-04 | NA | NA | 4.41E-10 | 1.49E-05 | 2.90E-03 | 3.72E-01 | 7.56E-02 | 2.75E-01 | 1.02E-05 | 3.12E-02 |
| | DLSNav | NA | NA | 7.94E-04 | 8.51E-02 | 2.27E-01 | 8.38E-03 | 9.70E-09 | 2.44E-04 | NA | NA | 4.41E-10 | 4.09E-07 | 1.23E-03 | 2.63E-01 | 1.22E-01 | 7.10E-02 | 4.10E-05 | 7.76E-03 |

A - P-values of unscaled L, N, and PLmax values between the AI datasets to the normal GC and PBL controls. B - P-values of PL and DLFSN values scaled by sequence alignment length in comparisons between the AI datasets to the normal GC and PBL controls. C - P-values of L and OD values when scaled by tree pick size in comparisons between the AI datasets to the normal GC and PBL controls. D - P-values of unscaled DASN and DLSN values between the AI datasets to the normal GC and PBL controls. E - values of DASN and DLSN values when scaled by sequence alignment length between the AI datasets to the normal GC and PBL controls. Highlighted values were found to be significant after FDR correction. Tree properties and dataset identities described in the text and in Supplementary Table 1.