Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing

Peter J. Campbell*, Erin D. Pleasance*, Philip J. Stephens*, Ed Dicks*, Richard Rance*, Ian Goodhead*, George A. Follows†, Anthony R. Green†, P. Andy Futreal*‡, and Michael R. Stratton*‡§

*Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom; †Department of Haematology, University of Cambridge, Cambridge CB2 2XY, United Kingdom; and §Institute of Cancer Research, Sutton, Surrey SW7 3RP, United Kingdom

Edited by Marshall Horwitz, University of Washington, Seattle, WA 98195, and accepted by the Editorial Board June 9, 2008 (received for review February 21, 2008)

During the clonal expansion of cancer from an ancestral cell with an initiating oncogenic mutation to symptomatic neoplasm, the occurrence of somatic mutations (both driver and passenger) can be used to track the on-going evolution of the neoplasm. All subclones within a cancer are phylogenetically related, with the prevalence of each subclone determined by its evolutionary fitness and the timing of its origin relative to other subclones. Recently developed massively parallel sequencing platforms promise the ability to detect rare subclones of genetic variants without *a priori* knowledge of the mutations involved. We used ultra-deep pyrosequencing to investigate intraclonal diversification at the Ig heavy chain locus in 22 patients with B-cell chronic lymphocytic leukemia. Analysis of a non-polymorphic control locus revealed artifactual insertions and deletions resulting from sequencing errors and base substitutions caused by polymerase misincorporation during PCR amplification. We developed an algorithm to differentiate genuine haplotypes of somatic hypermutations from such artifacts. This proved capable of detecting multiple rare subclones with frequencies as low as 1 in 5000 copies and allowed the characterization of phylogenetic interrelationships among subclones within each patient. This study demonstrates the potential for ultra-deep resequencing to recapitulate the dynamics of clonal evolution in cancer cell populations.

Cancer encapsulates many of the tenets underpinning evolutionary biology. Genetic variation within a cancer cell population, in the form of somatically acquired mutation, occurs randomly but at a defined rate influenced by, for example, exposure to carcinogens, frequency of cell turnover, and integrity of DNA repair pathways. Subclones of cells have a mixture of shared and private somatic changes that are subject to intense selective pressure and biological competition. The subclones with the best evolutionary fitness will, in time, come to dominate the cancer cell population, and each will be marked by the presence of mutations that provide a direct competitive advantage (driver mutations) and by others acquired during clonal evolution that contribute nothing to the subclone's oncogenic potential (passenger mutations) (1). Development of a symptomatic neoplasm may appear to be a linear procession at the macroscopic scale, but at the individual cellular level there will be many subclonal extinction events and coexistent, competing genotypes. Indirect experimental evidence for such models of cancer development include the high prevalence of somatically acquired mutations in some cancer types (1–3), the rapid emergence of preexisting drug resistant clones after selection with molecularly targeted therapy (4), and the long latency between initiating events and clinical presentation of a malignancy (5, 6). However, direct demonstration of multiple genetically related subclones within a tumor and their phylogenetic relationships has been hampered by the lack of tools for the detection of rare genetic variants.

Massively parallel sequencing has the potential to identify the compendium of rare subclones of genetic variants that may exist in human tumors in an unbiased fashion, unlike more conventional genotyping methods that require *a priori* knowledge of the mutations. This technique has been applied, for example, to the discovery of cells carrying drug resistance mutations before the initiation of therapy in both cancer (7) and infectious diseases (8, 9). To date, deep resequencing has detected variants down to a frequency of 1 in 100 (7–9), but its sensitivity for the detection of rarer variants has not been tested. With the appropriate informatic analyses and experimental design, the depth and breadth of sequencing available on the next-generation platforms will provide the tools to reconstruct clonal interrelationships of cancer cell populations, with relevance to identifying and tracking subpopulations of cells responsible for drug resistance, invasion, metastasis, and relapse, as well as annotating the genuinely initiating genetic lesions.

We sought to explore this potential using B-cell chronic lymphocytic leukemia (CLL) as a model system. It is a clonal malignancy of B lymphocytes in which each patient's leukemia carries a unique and functional VDJ rearrangement of the Ig heavy chain locus, *IGH*. More than half of CLL patients show evidence for extensive somatic hypermutation of a 300-bp region of the *IGH* rearrangement (10), using the same machinery that normal B cells enroll during the polyclonal response to antigen. There is growing evidence that interactions between autoantigens and the B-cell receptor, enhanced by certain patterns of hypermutation, are important in the clonal evolution and prognosis of chronic lymphocytic leukemia (11), although genomic lesions at other loci, such as microRNAs, also play a critical role in disease pathogenesis (12, 13). Recent reports have documented the existence in some CLL patients of subclones of leukemic cells that have patterns of somatic hypermutation within the *IGH* locus different from the dominant clone of cells (14–16), although the depth of these studies for probing the leukemic population is of necessity limited by the need to clone either PCR products or individual cells.

We therefore evaluated the potential of massively parallel pyrosequencing for identifying subclones of leukemic cells with variant *IGH* hypermutation in patients with CLL and for reconstructing the phylogenetic relationships. Although artifacts introduced during library preparation and sequencing become problematic for detecting variants with a frequency <1 in 100, sequence read-lengths of ≈ 250 bp from individual DNA molecules allowed us to identify haplotypes of somatic hypermutations carried by individual leukemic cells. This led to the characterization and quantitation of subclones as rare as 1 in 5000 and demonstrated the complexity of clonal evolution at the *IGH* locus in this disease.

Results

Deep Resequencing of the *IGH* Rearrangement in CLL. We developed a nested PCR approach to amplify the clone-specific *IGH* VDJ

Author contributions: P.J.C., G.A.F., A.R.G., P.A.F., and M.R.S. designed research; P.J.C., R.R., and I.G. performed research; P.J.C., E.D.P., P.J.S., E.D., A.R.G., P.A.F., and M.R.S. analyzed data; and P.J.C., E.D.P., P.J.S., G.A.F., A.R.G., P.A.F., and M.R.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.W. is a guest editor invited by the Editorial Board.

‡To whom correspondence may be addressed. E-mail: mrs@sanger.ac.uk or paf@sanger.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0801523105/DCSupplemental.

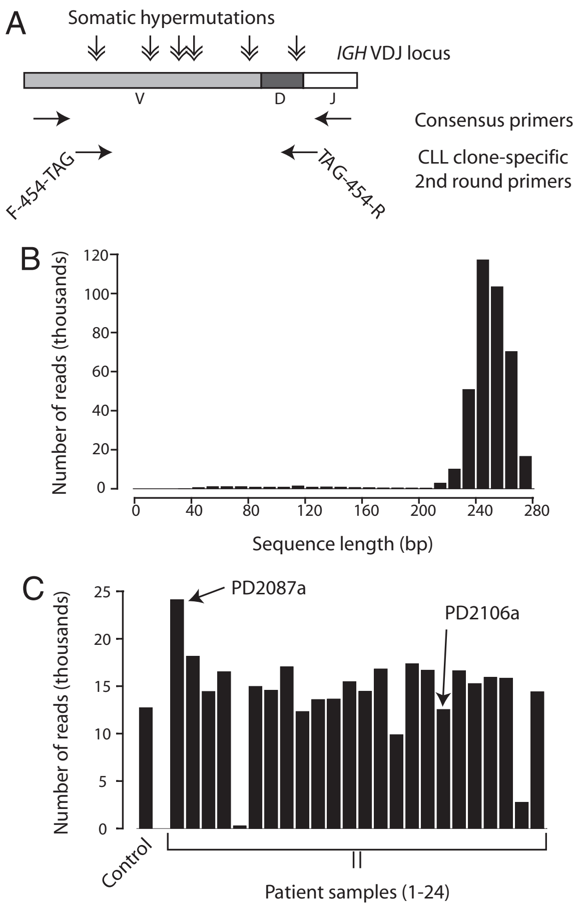

Fig. 1. Ultra-deep pyrosequencing of the *IGH* locus in patients with CLL. (*A*) Nested PCR to generate amplicons for resequencing used published consensus primers for the *IGH* locus in the first round. The second round of PCR used internal primers, specific to the patient's VDJ rearrangement, with a 5′ 3-bp barcode and forward (F-454) and reverse (R-454) sequencing primers. (*B*) Histogram of the sequence length of the ≈385,000 reads. (*C*) Histogram of the number of reads per patient, with the 2 sample codes (PD2087a and PD2106a) indicating the patients for whom phylogenetic trees were generated in Fig. 4.

rearrangement in 22 patients with B-cell CLL (Fig. 1*A*). The first round of amplification used published consensus primers for the *IGH* locus (17), and the second round used clone-specific nested primers. The PCRs from each patient were pooled and sequenced on the 454-FLX machine (Roche Diagnostics). Because each bead analyzed by pyrosequencing is effectively the sequence of a single DNA molecule, each read represents an *IGH* haplotype from a single leukemic cell.

A total of ≈385,000 sequences were generated using this protocol. Read length was remarkably consistent across the experiment (Fig. 1*B*). The average read length was 246 bp, indicating that the total yield from the sequencing run was ≈95 Mb. Using a nested PCR approach fulfilled 2 aims. First, the yield of PCR products was equalized across samples because the second round of PCR is not limited by template concentration and therefore always reaches saturation. Thus, except for 2 samples with disappointingly low yields, the number of sequences per patient was approximately comparable across the cohort (Fig. 1*C*). Second, because the primers were nested, generally only VDJ rearrangements from the CLL cells were amplified and sequenced. For every patient sample, <1% of the sequences generated were from contaminating polyclonal B cells.

Analysis of the Control Locus. To assess sequencing error rates, we included a control amplicon from the germ line *IGH* locus. The DNA used for the control sample came from normal renal tissue, was amplified using exactly the same PCR conditions as the CLL samples, and had no known polymorphisms in this region. We assumed that any variant base calls from the sequencing were artifacts.

Among incorrect base calls of apparently high quality, several distinct patterns of errors emerged (Fig. 2*A*). First, the pattern of insertions and deletions was not uniform across the locus, with certain positions of the amplicon being particularly susceptible to these artifacts. These errors were predominantly found around runs of 4 or more bases of the same nucleotide, known as "homopolymer tracts," and reflect the known difficulty of resolving these using pyrosequencing methods. Second, there were more errors of all types toward the end of the sequence than at the beginning. Third, a low rate of substitutions was found throughout the length of the amplicon.

We found 19255 artifactual insertions in the 3.33 Mb of sequence from the control locus, of which 5287 were high quality and recurrent (seen in ≥ 3 reads). The position of these insertions was predictable, with 4 patterns accounting for all but 14 of them [supporting information (SI) Fig. S1*A*]. During the pyrosequencing protocol, nucleotides are flow-released in a specific cyclical order (A, C, G, and then T). Around homopolymer tracts of a given base, an artifactual insertion of the same base often would be called at the flow-release of that nucleotide immediately before or after the repetitive tract. These 2 patterns accounted for 4201 and 115 of the insertions, respectively. This artifact probably represents phasing errors in the pyrosequencing in which individual molecules of DNA on the microbead may get ahead or lag behind the majority during the cyclical release of nucleotides. A third pattern, accounting for 29 of the insertions, was the insertion of the same base as in a homopolymer tract at the next true instance of that nucleotide after the tract. The final pattern, which accounted for 928 of the insertions, was insertion of nucleotides after 200 bp of the sequence had been read. This likely reflects the low signal-to-noise ratio seen toward the end of sequence reads. Of the 14 insertions that did not reflect any of these patterns, none was seen in both forward and reverse reads.

A total of 7822 artifactual deletions were called in the control sequences, with 2076 being high-quality and recurrent deletions. Many deletions occurred in primer sequence (497/2076, Fig. S1*B*) and therefore represent errors during primer synthesis rather than pyrosequencing artifacts. Apparent deletions at homopolymer tracts accounted for 1359 deletions. Similar to the patterns seen for insertions, we also saw artifactual deletions at the next instance of the same base after a homopolymer tract (76 deletions) and deletions after 200 bp of sequence had been read (133 of the 2076 deletions). Eleven deletions did not fall into any of these 4 patterns; all 11 occurred at the same base and exclusively in the reverse reads.

High-quality, recurrent substitutions tended to occur at a lower but more even rate across the amplicon and had the pattern of *Taq* polymerase errors. Examination of the pyrograms suggested they were unlikely to be pyrosequencing errors (Fig. S1*C*), because they were not necessarily associated with homopolymer tracts, often occurred early in the sequence read, had high signal-to-noise ratios, and were seen in both forward and reverse reads. Moreover, the observed substitution rate is higher than that quoted for high-quality substitutions caused by sequencing errors on the 454 platform (≈1–2/10 kb) (18). There was a definite bias in the base changes observed (Table 1). T>C; A>G transitions were the most frequent, with an observed high-quality error rate of 1.2 per 1000 sequenced bases. The next most common substitution was the other transition, G>A; C>T (0.37 per 1000 sequenced bases). Transversions were less common, all occurring at ≤ 0.15 per 1000 sequenced bases. The pattern of these base substitutions is similar to those observed in studies of PCR fidelity (19–21), suggesting that polymerase errors are the cause of most of the observed substitutions here, despite the use of a blend of *Taq* polymerase and proof-

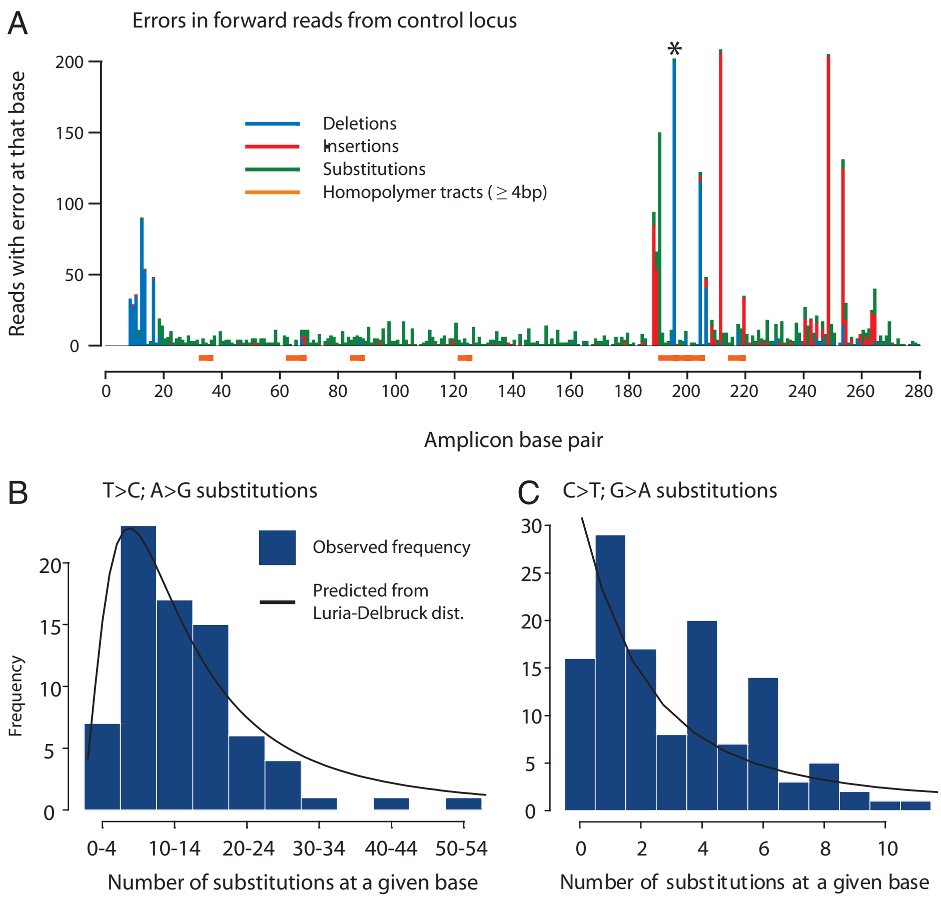

Fig. 2. Sequencing errors in the control locus. (*A*) Histogram showing the distribution and type of high-quality (Phred ≥ 20 for insertions and deletions; base height drop ≥ 0.75 peak height for deletions) sequencing errors in the forward reads from the control amplicon, together with the locations of homopolymer tracts (*orange*). The base marked with the asterisk actually had 922 deletions, beyond the range of the graph. (*B*) Observed (*bars*) and fitted theoretical (line) frequency distribution of T>C; A>G errors in the control locus. The number of substitutions at each T or A nucleotide >2 bases outside homopolymer tracts and primer sequence in the amplicon was counted across the 12,700 reads. The Luria–Delbrück distribution (line) was fitted to these observed frequencies, allowing estimation of the misincorporation rate of the polymerase. (*C*) Observed (*bars*) and fitted theoretical (*line*) frequency distribution of C>T; G>A errors in the control locus. The number of substitutions at each C or G nucleotide >2 bases outside homopolymer tracts and primer sequence in the amplicon was counted across the 12,700 reads. The Luria-Delbrück distribution (*line*) was fitted to these observed frequencies, allowing estimation of the misincorporation rate of the polymerase.

reading enzyme. The errors probably represent polymerase substitutions during library preparation rather than emulsion PCR, because the pyrosequencing trace suggests all molecules on the bead have the same substitution (Fig. S1*C*).

The rate of polymerase errors determines the fundamental limit of the ability of deep resequencing to detect non-artifactual single-base substitutions in PCR amplicons. We hypothesized that the biological properties of polymerase substitutions would be analogous to drug-resistance mutation rates in microbes, because such mutations also arise spontaneously in a population and double in size with each generation/cycle. The prevalence of microbial mutations follows the Luria-Delbrück distribution (22). We therefore fitted this model to the observed distribution of each of the 6 classes of base substitutions in our data. We found good concordance between the observed and fitted frequencies for the T>C; A>G transition (Fig. 2*B*) and for the C>T; G>A transition (Fig. 2*C*), as well as the transversions (data not shown). This concordance provides further support for the hypothesis that the observed substitutions do indeed arise from polymerase errors, and the predicted distribution also allows the calculation of the probability that a given number of substitutions at a particular base could have arisen as a consequence of *Taq* misincorporation.

An Algorithm to Detect Rare Subclones. The challenge for any deep-resequencing experiment is to distinguish the rare genuine variants from artifacts generated during library preparation or sequencing. We developed a bioinformatic algorithm based on the preceding analysis of the control locus to identify reads containing genuine variants (Fig. S2). Artifacts caused by pyrosequencing errors were eliminated by ignoring any apparent substitutions, insertions, or deletions occurring near or in homopolymer tracts and requiring any pattern of variants to be seen in both forward and reverse reads. To eliminate artifactual substitutions caused by *Taq* polymerase errors, we used our observation that they are random events, occurring independently at a fixed rate for each particular base change. In contrast, genuine somatic hypermutations in subclones of CLL cells often show several substitutions linked on the same DNA molecule, namely haplotypes of variants. Before accepting a haplotype of variants as a genuine subclone, we required it to be seen at least 3 times, to be seen in both forward and reverse reads, and to show no evidence of poor-quality sequence on manual examination of the pyrosequencing trace.

Analysis of Patient Samples. The algorithm outlined in the previous sections was applied to the sequencing data from patient samples and did indeed identify rare haplotypes showing a pattern of somatic hypermutation different from that seen in the dominant clone. In fact, we were able to detect subclones down to a frequency of 1 in 5000 reads, confirming the exquisite sensitivity of the deep resequencing approach. Examination of the pyrosequencing traces confirmed that these variant mutations were found in high-quality sequence reads, away from homopolymer tracts and often in close linkage with each other (Fig. 3*A*). Furthermore, as exemplified in Fig. 3*A*, different subclones often shared mutations that were not found in the dominant clone, suggesting they were clonally related.

Six of the 24 samples (25%) showed evidence for >1 clone in their population of CLL cells, with the number of identified clones per sample varying from 2 to 18 (Fig. 3*B* and Table S1). Of the 16 patients who had somatic hypermutation of >2% of bases in their dominant *IGH* clone, 4 showed evidence for intraclonal diversification, and 1 of the patients who had both *IGH* alleles rearranged showed evidence for intraclonal diversification in both alleles. Surprisingly, however, 1 of the patients with 3 subclones (PD2099a; Table S1) actually had no mutations in the dominant clone of CLL

Table 1. Rate of high quality (Phred score ≥ 20) artifactual nucleotide substitutions at least 2 bp outside homopolymer tracts and excluding primer sequences observed in 12,700 reads from the control locus

Substitution	Number*	Number bases sequenced*	Observed error rate
T → C A → G	1073	852,624	1.2×10^{-3}
G → A C → T	504	1,373,928	3.7×10^{-4}
T → G A → C	128	852,624	1.5×10^{-4}
G → C C → G	174	1,373,928	1.3×10^{-4}
T → A A → T	71	852,624	8.3×10^{-5}
G → T C → A	85	1,373,928	6.2×10^{-5}
Total	2035	2,226,552	9×10^{-4}

*Excluding bases sequenced within 2 bp either side of a homopolymer tract and primer sequence.

cells, suggesting that the phenomenon of clonal evolution of the VDJ rearrangement is not restricted to the subgroup of patients with established somatic hypermutation. The prevalence of individual subclones varied over several logs, ranging from 0.09 of the total population for the most frequent down to the limit of detection for the depth of sequencing in this study, 0.0002. There were no differences in clinical features between patients who had evidence for intraclonal diversification and those who did not (Table S2), although the sample size is too small for definitive conclusions.

Many of the identified subclones showed a surprising number of bases that were different from the dominant clone (Fig. 3*C*), ranging up to 18 variant bases, even though the amplicon length was <300 bp. However, approximately half of the subclones had 5 or fewer variant bases compared with the dominant clone, and this may be underestimated given that genuine subclones with a 1-base difference are difficult to distinguish from *Taq* errors.

Several features of the subclones identified in this study argue strongly that the subclones arise from the somatic hypermutation machinery and therefore represent genuine stages in the clonal evolution of CLL in these patients. Most significantly, many of the differences seen between individual subclones and the dominant clone actually represent the germ line base in the subclone and a mutated base in the dominant clone (seen in 15 of the 17 subclones from PD2087a and all 6 from PD2106a). Up to 7 bases in some subclones showed this pattern, a number much too large to be explained by artifactual errors of library preparation or pyrosequencing. Furthermore, for each of the 5 samples with somatic hypermutation of the IGH locus in the dominant clone, the subclones and dominant clone shared a core subset of mutations, confirming the clonal relationship between the subclones and the dominant clone of cells. Furthermore, the subclonal mutations showed many of the properties of somatic hypermutation. For example, the transition:transversion ratio was ≈ 1:1, as expected. No genuine insertions or deletions were identified, which is not surprising because they often cause frameshifts. Moreover, the subclonal mutations falling in the structural regions of the heavy chain, known as the "framework regions," showed that the rate of replacement mutations (those causing an amino acid substitution) was significantly less than would be expected by chance (PD2087a Replacement:Silent (R:S) ratio 12:12, $P = 0.0008$; PD2106a 1:6, $P = 0.01$), whereas mutations in the complementarity-determining regions showed a pattern more consistent with that expected by chance (PD2087a R:S ratio 12:6, $P = 0.2$; PD2106a R:S ratio 1:1, $P = 0.8$). This difference between R:S ratios in the structural and antigen-binding regions of the *IGH* locus is a well recognized property of somatically hypermutated VDJ rearrangements under selection pressure (14, 23).

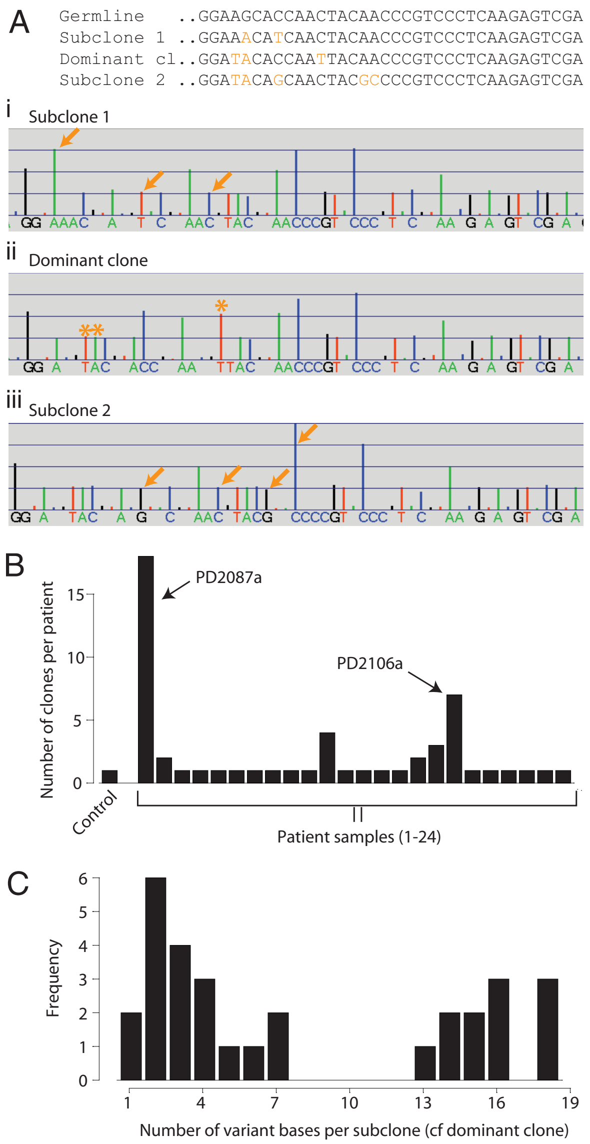

Fig. 3. Identification of rare subclones at the *IGH* locus in patients with CLL. (*A*) Pyrosequencing traces from the dominant clone (*ii*) and 2 related subclones (*i* and *iii*) in a patient with CLL show closely linked base substitutions compared with the dominant clone: T>A; C>T; T>C in subclone 1 and C>G; T>C; A>G; A>C in subclone 2. The dominant clone has several mutations compared with the germ line, marked with an asterisk, and several of the variants in the subclones are the germ line sequence (e.g., T>A and T>C in subclone 1). The 2 subclones share the T>C variant, suggesting their clonal relationship. (*B*) Histogram showing the number of observed clones per sample, with the 2 sample codes (PD2087a and PD2106a) indicating the patients for whom phylogenetic trees were generated in Fig. 4. (*C*) Histogram showing the number of bases different between each subclone and the dominant clone for that patient.

Reconstructing the Phylogenetic Relationships of Leukemic Subclones. These observations suggest that the evolutionary relationships among the subclones and dominant clone of CLL cells may provide insights into the clonal development of the leukemia. We therefore fitted unrooted parsimony models to generate phylogenetic trees for the 2 samples with the most complex and numerous subclones, PD2087a (Fig. 4*A*) and PD2106a (Fig. 4*B*).

Bootstrapping was performed to evaluate the reproducibility of the trees. This demonstrated that although exact ancestral relation-

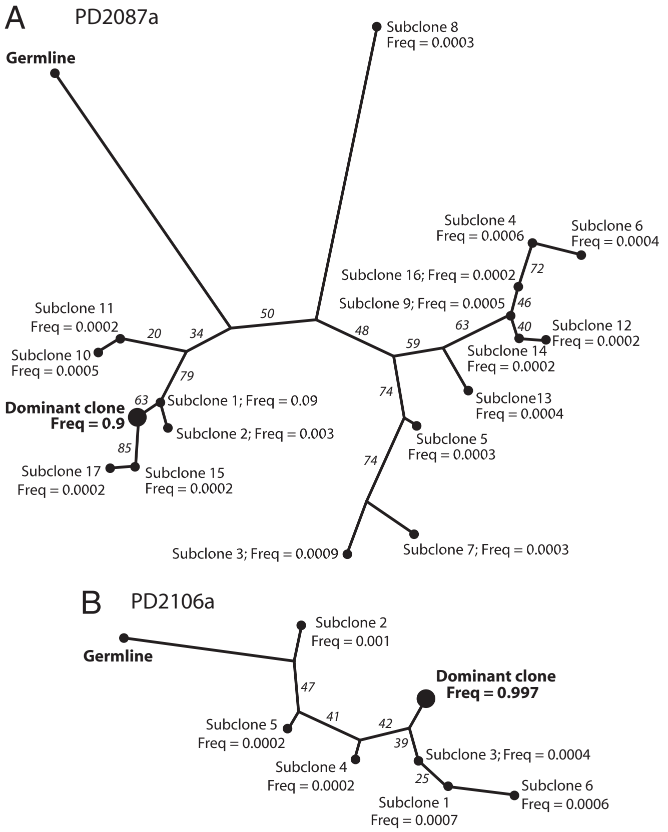

Fig. 4. Trees showing the phylogenetic interrelationships among the subclones for 2 patients, (*A*) PD2087a and (*B*) PD2106a. The trees were fitted using unrooted parsimony methods, and the length of each branch is proportional to the number of varying bases (evolutionary distance). The number shown beside each intermediate branch is the percentage support across 1000 bootstrap samples.

ships among individual subclones show some uncertainty, there is strong support for 3 different classes of subclones: intermediate stages on the route to the dominant clone, blind alleys representing divergent evolution from a common ancestor, and on-going evolution from the dominant clone itself. The persistence of the intermediate stages and divergent subclones suggests that the initiating driver mutation(s) in leukemogenesis must have occurred at or before the earliest branch-point of the tree. That the dominant clone shows further evolution from this ancestral branch-point suggests that at least 1 other driver mutation must have occurred to give it a selective advantage over the other subclones; otherwise the dominant clone could not have become so numerically dominant. There is extensive evidence that the *IGH*-locus hypermutations themselves play a driver role in these clonal expansions (11). Here we observed that mutations shared by all clones (i.e., the earliest events) showed a somewhat higher rate of replacement mutations (R:S ratio 10:4) compared with mutations seen only in some subclones (R:S 26:25). Although greater numbers of patients need to be studied to confirm the pattern, this finding would be consistent with the hypothesis that somatic hypermutation of the antigen receptor plays an early role in leukemia development (11) and that later mutations (i.e., those evident only in some subclones) are subject to negative selection pressure if they cause significant alterations to protein structure.

Discussion

This study illustrates the informatic analyses required to use ultra-deep resequencing for the detection of rare DNA variants at a particular locus. Previous studies have demonstrated sensitivity down to 1 in 100 copies (7–9), and we have extended this to detect genuine haplotypes at the *IGH* locus in B-cell CLL that are as rare as 1 in 5000 copies without *a priori* knowledge of their sequence. The wealth of detail yielded by the sequencing allowed the documentation of phylogenetic interrelationships among the subclones to a much greater resolution than previous reports of intraclonal diversification in CLL (14–16). In particular, we have identified rarer subclones than possible with more conventional methods, have found evidence that at least 2 driver mutations underlie the clonal expansions seen in 2 patients with CLL, and have provided support for the hypothesis that positive selection for somatic hypermutation of the antigen receptor is an early event during leukemogenesis (11).

We have demonstrated that polymerase errors determine the maximum sensitivity of deep resequencing for the detection of single-base changes in an amplicon. This concept is particularly important because many potential applications of deep resequencing involve the detection of single amino acid substitutions, such as imatinib resistance mutations in chronic myeloid leukemia (24) and gefitinib resistance mutations in lung cancer (7, 25). For our experiment, for example, with the nested PCR protocol and the blend of *Taq* polymerase and proof-reading enzyme we used, we would be unable to designate a single T-to-C transition as genuine unless it was present at $>$ 1 in 50 copies, or a T-to-A transversion as genuine unless it was present in $>$ 1 in 1000 copies. This is even more critical for studies using cDNA, because reverse transcriptases have error rates at least a log higher than DNA polymerases (26).

These data have several implications for our understanding of the clonal evolution of CLL and extend previous reports of these subclones (14–16). The preponderance of silent mutations in the structural regions of the *IGH* gene in many of the subclones suggests that significant negative selection pressure has acted at this locus during the intraclonal diversification of the leukemia, eliminating deleterious replacement mutations, as described previously (14). Unexpectedly, subclones are not restricted to patients with extensive hypermutation in their dominant clone. Not all patients showed evidence for multiple *IGH* subclones, however. It is possible that intraclonal diversification in these samples was missed, either because the PCR amplification and sequencing failed to sample all subclones or because the subclones were too rare to detect at the depth of sequencing performed in this study.

Many of the observed subclones represent intermediate stages or divergent branches from the route of germ line configuration to dominant clone (Fig. 4). Because a patient has many billions of circulating CLL cells at any given time, and this population turns over at a rate as high as 1% per day (27), even a subclone with a frequency of 1 in 5000 must have significant leukemogenic potential in its own right. Therefore, we can hypothesize that all these subclones share at least 1 initiating genetic lesion with oncogenic potential. Furthermore, for the dominant clone to become numerically predominant, it must have a selective advantage over its direct predecessors, suggesting the existence of at least 1 more driver mutation in that clone, whether it be in the *IGH* locus or elsewhere.

This study demonstrates the potential of new sequencing technologies for the delineation of clonal evolution in cancer. Although CLL is a particularly strong model system for testing these concepts, because of the high mutation rate in a stretch of DNA that is sufficiently short to be covered with a single sequencing read, it certainly is possible to conceive similar studies in other cancers. Phylogenetic trees could be established by genotyping distinct populations of cancer cells, whether they are sorted individual cells or samples from geographically or temporally distinct parts of the tumor, including regions of invasion, metastasis, or relapse.

Methods

Patient Samples. Patients were recruited from the CLL clinic at Addenbrooke's National Health Service Foundation Trust, Cambridge, U.K., after giving written informed consent. The study had approval from the local research ethics committee. All patients had B-cell CLL diagnosed using standard criteria on the basis of characteristic peripheral blood morphology and CD5+, CD19+, monoclonal light-chain immunophenotype. We studied anonymous DNA samples from unsorted mononuclear cells of 22 patients, 16 of whom had > 2% somatic hypermutation of the *IGH* locus. In 2 patients both *IGH* loci were rearranged; in both cases a frameshift mutation in 1 of the alleles rendered it nonfunctional. For these 2 samples, both rearrangements were investigated with deep resequencing. In addition to the 24 patient amplicons, a control locus within the germ line configuration of the *IGH* locus using DNA from a normal renal biopsy was sequenced.

454 Sequencing. Six PCRs, representing each of the 6 families of V_H segments, for each patient's VDJ rearrangement were performed using the BIOMED consensus forward primers, all binding in framework region 1, and the consensus reverse primer for the J_H segment (17). The products were visualized on an agarose gel, and the monoclonal band from the leukemic cells was excised and Sanger sequenced. Nested primers were designed in which the 5′ sequence of the primer was specific for the 454-FLX emulsion PCR, followed by a 3-bp barcode unique to each patient, and completed by an ≈20-bp primer sequence specific to the given patient's *IGH* rearrangement. PCRs were performed using a blend of *Taq* polymerase and proof-reading enzyme (High fidelity Platinum *Taq*; Invitrogen), with 35 cycles for each round.

Amplicons from the patient and control samples were purified and pooled before performing emulsion PCR, using the same polymerase/proof-reading blend. Two emulsion PCRs were performed, for forward and reverse sequencing, and microbeads then were collected, combined, and deposited onto slides provided for the 454-FLX instrument (Roche Diagnostics). Massively parallel pyrosequencing was performed according to the manufacturer's protocol, and base calls and quality scores were extracted using the GS-amplicon software provided with the platform.

Bioinformatic and Statistical Analysis. A flow-chart of the bioinformatic algorithm used to identify genuine haplotypes of variants is shown in Fig. S2. Sequences derived from each patient were identified from their unique 3-bp barcode and aligned with the dominant clone for that patient using the BLAST algorithm (BLASTN v2.2.15). Insertions and substitutions (with their corresponding Phred quality scores) were extracted, as were deletions (with the associated fall in peak height) for each sequence. Homopolymer tracts, defined as ≥4 bp of the same base in a row, were identified in the dominant clone's sequence, and any insertions occurring at the flow release of that nucleotide immediately before or after the tract or at the next instance of that base after a tract were excluded from further analysis. Low-quality insertions (quality score < 20) were excluded also. Only deletions with a drop of ≥ 0.75 of the base's height were considered. Deletions at homopolymer tracts were excluded.

To identify *Taq* error rates, the Luria-Delbrück distribution was fitted to the number of high-quality (≥ 20) substitutions of a given type (for example, T>C; A>G) occurring at each at-risk base (i.e., each T or A) along the amplicon. The observed median was calculated using the Jones estimator, as described (22) and was used as the basis for calculating the maximum likelihood of the mutation rate via the recursive MSS algorithm (22). With this estimate of the mutation rate and the known number of reads covering a given base as the denominator, we could calculate the expected frequency distribution for the number of substitutions at this base. Reads carrying only a single substitution compared with the dominant clone were counted as genuine if the frequency of the given substitution exceeded the 98% threshold for *Taq* errors thus calculated. To calculate the appropriate threshold for reads with 2 substitutions compared with the dominant clone, we inputted the number of reads with just 1 of the 2 substitutions as the denominator in the same calculation. The chances of 3 polymerase errors clustering on the same DNA molecule (maximum frequency 1/500,000 per amplicon) was substantially less than the depth to which we sequenced the amplicons, and these clusters therefore were considered genuine.

Thus, an algorithm to identify haplotypes of variants was developed. Unique haplotypes of high-quality substitutions (quality ≥ 20) were identified in the dataset of sequences for a given sample. Where there were only 1 or 2 variants compared with the dominant clone, they would be considered genuine only if of sufficient frequency to exceed the expected number of polymerase errors. Only haplotypes seen in both forward and reverse reads were considered genuine to exclude strand- and context-specific sequencing errors. Once variant reads had been identified, these were examined manually to confirm that observed mutations were of good quality and convincing and to check whether mutations seen in reads from other subclones were present (but at lower quality than the 20 threshold). This latter step was important because occasionally, especially toward the end of sequence reads, genuine hypermutations found in other subclones might be missed.

The sequences of clones and subclones for a given patient underwent multiple alignment using the ClustalW2 algorithm (www.ebi.ac.uk/Tools/clustalw2/index.html) with default parameters and phylogenetic trees fitted using unrooted parsimony methods implemented in Phylip (http://evolution.gs.washington.edu/phylip.html) with default parameters. Reproducibility of the phylogenetic trees was assessed using 1000 bootstrap samples of individual nucleotides in the multiple alignment. The statistical analysis of replacement and silent mutations in framework and complementarity-determining regions was performed using multinomial distribution methods, as described (23), on the authors' applet (http://www-stat.stanford.edu/lg/).

ACKNOWLEDGMENTS. P.J.C. is funded by the Kay Kendall Leukaemia Fund. The Cancer Genome Project is supported by the Wellcome Trust. We acknowledge the assistance of the Addenbrooke's Hospital Haematological Disorders Sample Bank in processing and providing the samples for analysis.

1. Greenman C, *et al.* (2007) Patterns of somatic mutation in human cancer Genomes. *Nature* 446:153–158.
2. Sjoblom T, *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–274.
3. Davies H, *et al.* (2005) Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* 65:7591–7595.
4. Roche-Lestienne C, *et al.* (2002) Several types of mutations of the *Abl* gene can be found in chronic myeloid leukemia patients resistant to STI571, and they can pre-exist to the onset of treatment. *Blood* 100:1014–1018.
5. Greaves MF, Wiemels J (2003) Origins of chromosome translocations in childhood leukemia. *Nat Rev Cancer* 3:639–649.
6. Hong D, *et al.* (2008) Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science* 319:336–339.
7. Thomas RK, *et al.* (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* 12:852–855.
8. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Res* 17:1195–1201.
9. Hoffmann C, *et al.* (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 35:e91.
10. Messmer BT, Albesiano E, Messmer D, Chiorazzi N (2004) The pattern and distribution of immunoglobulin VH gene mutations in chronic lymphocytic leukemia B cells are consistent with the canonical somatic hypermutation process. *Blood* 103:3490–3495.
11. Ghia P, Caligaris-Cappio F (2006) The origin of B-cell chronic lymphocytic leukemia. *Semin Oncol* 33:150–156.
12. Cimmino A, *et al.* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci USA* 102:13944–13949.
13. Calin GA. *et al.* (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* 353:1793–1801.
14. Volkheimer AD, *et al.* (2007) Progressive immunoglobulin gene mutations in chronic lymphocytic leukemia: Evidence for antigen-driven intraclonal diversification. *Blood* 109:1559–1567.
15. Bagnara D, *et al.* (2006) IgV gene intraclonal diversification and clonal evolution in B-cell chronic lymphocytic leukaemia. *Br J Haematol* 133:50–58.
16. Gurrieri C, *et al.* (2002) Chronic lymphocytic leukemia B cells can undergo somatic hypermutation and intraclonal immunoglobulin V(H)DJ(H) gene diversification. *J Exp Med* 196:629–639.
17. van Dongen JJ, *et al.* (2003) Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: Report of the BIOMED-2 Concerted Action BMH4-CT98–3936. *Leukemia* 17:2257–2317.
18. Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: An improved base caller for SNP discovery in pyrosequences. *Nat Methods* 5:179–181.
19. Bracho MA, Moya A, Barrio E (1998) Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. *J Gen Virol* 79 (Pt 12):2921–2928.
20. Dunning AM, Talmud P, Humphries SE (1988) Errors in the polymerase chain reaction. *Nucleic Acids Res* 16:10393.
21. Ennis PD, Zemmour J, Salter RD, Parham P (1990) Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: Frequency and nature of errors produced in amplification. *Proc Natl Acad Sci USA* 87:2833–2837.
22. Foster PL (2006) Methods for determining spontaneous mutation rates. *Methods Enzymol* 409:195–213.
23. Lossos IS, Tibshirani R, Narasimhan B, Levy R (2000) The inference of antigen selection on Ig genes. *J Immunol* 165:5122–5126.
24. O'Hare T, Eide CA, Deininger MW (2007) Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood* 110:2242–2249.
25. Kobayashi S, *et al.* (2005) EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *N Engl J Med* 352:786–792.
26. Arezi B, Hogrefe HH (2007) Escherichia coli DNA polymerase III epsilon subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures. *Anal Biochem* 360:84–91.
27. Messmer BT, *et al.* (2005) In vivo measurements document the dynamic cellular kinetics of chronic lymphocytic leukemia B cells. *J Clin Invest* 115:755–764.