novoSNP, a novel computational tool for sequence variation discovery

Stefan Weckx, Jurgen Del-Favero, Rosa Rademakers, et al.

Genome Res. 2005 15: 436-442
Access the most recent version at doi:10.1101/gr.2754005

Resource

novoSNP, a novel computational tool for sequence variation discovery

Stefan Weckx, Jurgen Del-Favero,[1] Rosa Rademakers, Lieve Claes, Marc Cruts, Peter De Jonghe, Christine Van Broeckhoven, and Peter De Rijk

Department of Molecular Genetics, Flanders Interuniversity Institute for Biotechnology, University of Antwerp, Antwerpen, Belgium

Technological improvements shifted sequencing from low-throughput, work-intensive, gel-based systems to high-throughput capillary systems. This resulted in a broad use of genomic resequencing to identify sequence variations in genes and regulatory, as well as extended genomic regions. We describe a software package, novoSNP, that conscientiously discovers single nucleotide polymorphisms (SNPs) and insertion–deletion polymorphisms (INDELs) in sequence trace files in a fast, reliable, and user-friendly way. We compared the performance of novoSNP with that of PolyPhred and PolyBayes on two data sets. The first data set comprised 1028 sequence trace files obtained from diagnostic mutation analyses of *SCNIA* (neuronal voltage-gated sodium channel α-subunit type I gene). The second data set comprised 9062 sequence trace files from a genomic resequencing project aiming at the construction of a high-density SNP map of *MAPT* (microtubule-associated protein tau gene). Visual inspection of these data sets had identified 38 sequence variations for *SCNIA* and 488 for *MAPT*. novoSNP automatically identified all 38 *SCNIA* variations including five INDELs, while for *MAPT* only 15 of the 488 variations were not correctly marked. PolyPhred detected far fewer SNPs as compared to novoSNP and missed nearly all INDELs. PolyBayes, designed for the sequence analysis of cloned templates, detected only a limited number of the variations present in the data set. Besides the significant improvement in the automated detection of sequence variations both in diagnostic mutation analyses and in SNP discovery projects, novoSNP also offers a user-friendly interface for inspecting possible genetic variations.

[novoSNP is freely available online at http://www.molgen.ua.ac.be/bioinfo/novosnp.]

With the human and numerous other eukaryotic genome sequences finished (The *C. elegans* Sequencing Consortium 1998; Adams et al. 2000; Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002), and other genome-wide sequencing efforts ongoing, researchers can now fully explore these genomes. Mapping genetic differences between individuals is one of the major challenges in the post-genome era, which will provide valuable information about the quality of life of human beings. Discovery of these sequence variations by resequencing of a genomic region in a set of individuals is considered the golden standard (Kwok et al. 1994).

Those sequence variations are mostly mutations in coding or regulatory regions of transcription units, potentially related to a phenotypic trait or disease, or single nucleotide polymorphisms (SNPs) that can be used as markers for genetic associations studies, for fine mapping candidate regions based on linkage disequilibrium, or in pharmacogenetics aiming at genetic profiling patients for drug response and/or side effects. Since SNPs occur on average once every 1000 bp (Sachidanandam et al. 2001), they offer a higher marker density compared to short tandem repeat (STR) markers and allow high-throughput automated analysis. Therefore, SNPs are rapidly becoming the genetic markers of choice, especially in the search for genetic factors involved in complex diseases or traits, and in pharmacogenetics. A high-density SNP map of a gene or a chromosomal candidate region can be constructed using data from public SNP databases (Sherry et al. 2001; Fredman et al. 2004). However, since the number of validated SNPs is often still limited and marker density is not always sufficiently high, additional sequencing efforts are often needed to saturate a gene or candidate chromosomal region with SNPs.

As sequencing has evolved from low-throughput, work-intensive, gel-based systems to high-throughput capillary systems, data analysis is becoming a major bottleneck in a resequencing approach. Several sequence-variation-finding programs like PolyPhred (Nickerson et al. 1997) and PolyBayes (Marth et al. 1999) are available. However, these programs have shown limitations regarding correct SNP and/or INDEL discovery. Therefore, we developed novoSNP, providing a fast, reliable, and accurate strategy for the discovery of SNPs and INDELs from sequence trace files obtained from large-scale genomic resequencing projects of genes or candidate chromosomal regions.

Results

novoSNP-based automated sequence variation discovery is a straightforward process that can be divided into two major steps: detection and validation. Both steps are supported from within an intuitive graphical user interface.

Initiating a project

The first step in initiating a new novoSNP project is the creation of a single file SQLite database and the addition of a reference sequence in FASTA format. Next, one or more sets of sequence trace files can be added. Each set consists of forward and/or reverse sequence trace files from a region enclosed in the reference

[1]**Corresponding author.**
E-mail jurgen.delfavero@ua.ac.be; fax 32 3 820 2541.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.2754005.

15:436–442 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

sequence. Adding sequence trace files to the project automatically initiates SNP and INDEL detection according to the strategy described in the Methods section. All resulting scores are stored in the SQLite database.

The graphical user interface at a glance

The graphical user interface supports the validation step and consists of three frames and a toolbar with buttons for the most frequently used functions (Fig. 1). The sequence traces are visualized in the main frame (Fig. 1A). The reference sequence and the list of all sequence traces in the project are shown in a second frame (Fig. 1B). All identified sequence variations can be retrieved from the database and are presented in the variation display list (Fig. 1C). By clicking on a sequence variation from this list, the sequence trace views are shown in the main window with the variation highlighted in the center of the frame (Fig. 1A). To allow efficient validation of sequencing data, the software clusters similar sequence traces using a greedy algorithm, displaying one sequence trace for each group (Fig. 1D). Other sequence traces from within a group can be displayed by selecting the trace name in the list next to the displayed read (Fig. 1D). Parameter setting for sequence trace grouping can be adapted, and setting it to zero will display all traces separately.

The variation display list can be filtered and sorted in several ways: by variation type, by minimal and maximal overall score, by status of validation, and by start and end base position (Fig. 1E). During visual inspection, sequence variations can be annotated as approved, rejected, or uncertain with instant updating of the database (Fig. 1C).

To view the alignment of all sequences, an alignment window can be opened. The columns are color-coded according to the variation score, ranging from a white background representing a low score to red for a high overall score. At any time during the validation process, it is possible to exclude or include one or more sequence trace files and to reanalyze the remaining traces. Also, the structure of the file name can be adapted, allowing novoSNP to determine which forward and reverse reads originated from the same DNA sample. The variation data can be exported in text format, ordered by position or by sequence trace filename.

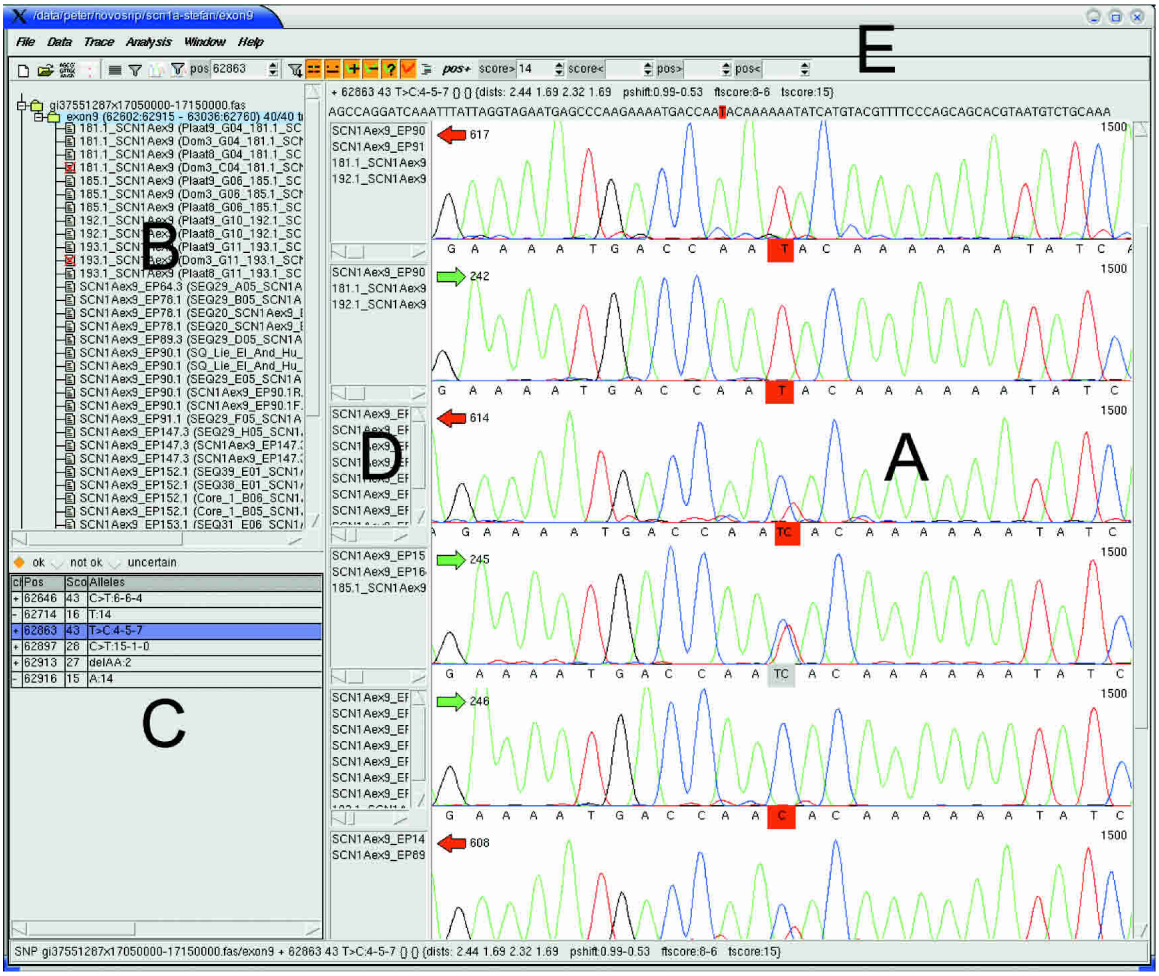

Figure 1. The graphical user interface. (*A*) The main frame, displaying trace files centered on a T/C SNP. (*B*) Window displaying an overview of the files in this project. (*C*) Window displaying the potential variations. The three checkboxes indicate whether a variation is approved (ok, +), rejected (not ok, −), or uncertain (?). (*D*) Sequence file names of the clustered sequence trace files. (*E*) The multifunctional toolbar.

Study of the performance of novoSNP, PolyPhred, and PolyBayes

The performance of novoSNP, PolyPhred, and PolyBayes was compared on two data sets, representing the two main resequencing approaches. A large data set, containing 9062 sequence trace files covering a 140-kb genomic region—the *MAPT* data set—illustrates a large-scale SNP discovery project with the typical tradeoff between throughput and the number of identified variations. A smaller data set, containing 1028 sequence trace files—the *SCN1A* data set—represents a typical gene mutation analysis project, requiring an extensive optimization of the resequencing process in order to ensure detection of all variations (Claes et al. 2003; Rademakers et al. 2004).

All three programs provide a quality score for each detected variation. Depending on the quality score cutoff used, several SNPs are detected for each program (Table 1). At the lowest-quality cutoff score, novoSNP detected all 38 variations in the *SCN1A* data set that were previously observed by visual inspection, including five INDELs, and missed only 10 out of 452 known SNPs (2.2%) and five out of 36 INDELs in the *MAPT* data set (Table 1). PolyPhred found all but three of the SNPs in the *SCN1A* data set at the lowest cutoff, but missed all five INDELs (Table 1A) while listing more false-positive INDELs (23) than novoSNP (nine). PolyPhred analysis of the *MAPT* data set showed that a large number of SNPs (172, or 38.1%) were not detected (Table 1B) and also that only two of 36 INDELs were correctly identified, while the number of false-positive INDELs (101) was again higher compared to novoSNP (63). PolyBayes was included in this comparative analysis as it is often used for SNP discovery. However, it was designed to handle sequencing data generated from cloned DNA templates (Marth et al. 1999) and is therefore unable to detect heterozygous bases and/or INDELs. Because of these limitations, PolyBayes identified only a small percentage of the SNPs in the *SCN1A* data set (54.5%) and the *MAPT* data set (31%) (Table 1). An overall comparison of the true SNPs and false positives (FP) detected by the three programs is represented as a Venn diagram in Figure 2. Clearly, novoSNP detected most SNPs for both data sets. However, PolyPhred detected three of the 10 SNPs missed by novoSNP. Somewhat surprisingly, most of the false positives were not shared between the different programs but were program-specific.

The use of low-quality cutoff values resulted in a large number of false positives for all three programs (Table 1). Using higher-quality cutoffs, at the expense of detecting less true variations, diminished the number of false positives. Only a small number of false positives remained when novoSNP was used with a high-quality cutoff of 20, while PolyPhred returned a substantially larger number of false positives (ranging from a factor 10 to 100 compared to novoSNP) with the highest-quality cutoff of 99

Table 1. Output summary of the novoSNP, PolyPhred, and PolyBayes SNP analysis on the *SCN1A* mutation and *MAPT* SNP data sets analyzed under different quality cutoff values

	Quality cutoff	Total number of SNPs	Correctly identified	False positives		False negatives	
A. *SCN1A*							
novoSNP	10	447	33	414	92.6%	0	0.0%
	15	122	32	90	73.8%	1	3.0%
	20	36	26	10	27.8%	7	21.2%
	25	26	22	4	15.4%	11	33.3%
PolyPhred	20	586	30	556	94.9%	3	9.1%
	25	510	30	480	94.1%	3	9.1%
	50	347	30	317	91.4%	3	9.1%
	75	254	30	224	88.2%	3	9.1%
	95	208	30	178	85.6%	3	9.1%
	99	189	26	163	86.2%	7	21.2%
PolyBayes	0.1	54	18	36	66.7%	15	45.5%
	0.25	46	17	29	63.0%	16	48.5%
	0.5	37	16	21	56.8%	17	51.5%
	0.75	33	16	17	51.5%	17	51.5%
B. *MAPT*							
novoSNP	5	3424	442	2982	87.1%	10	2.2%
	10	1146	421	725	63.3%	31	6.9%
	15	484	377	107	22.1%	75	16.6%
	20	251	244	7	2.8%	208	46.0%
	25	206	203	3	1.5%	249	55.1%
PolyPhred	20	2637	280	2357	89.4%	172	38.1%
	25	2510	280	2230	88.8%	172	38.1%
	50	2243	271	1972	87.9%	181	40.0%
	75	1892	252	1640	86.7%	200	44.2%
	95	1677	207	1470	87.7%	245	54.2%
	99	1572	175	1397	88.9%	277	61.3%
PolyBayes	0.1	991	140	851	85.9%	312	69.0%
	0.25	830	136	694	83.6%	316	69.9%
	0.5	672	126	546	81.2%	326	72.1%
	0.75	567	115	452	79.7%	337	74.6%

For the *SCN1A* data set, the lowest novoSNP shown cutoff is 10 since all SNPs were found at this cutoff value.

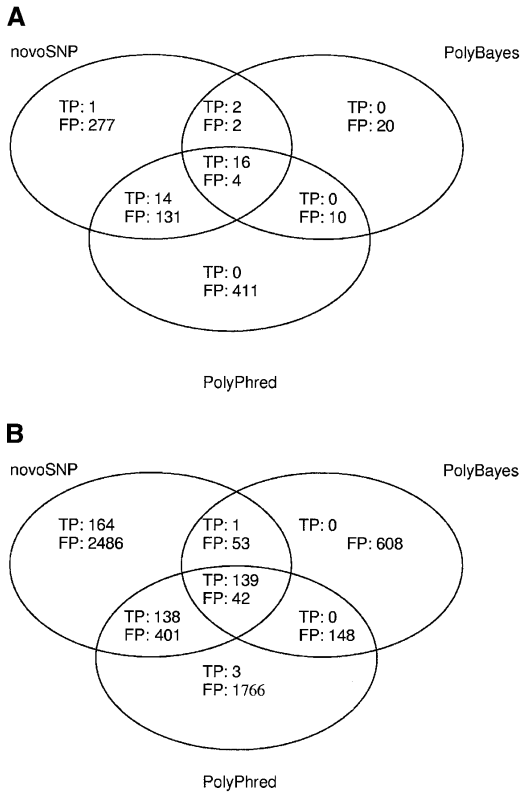

Figure 2. Venn diagrams representing the results of novoSNP, PolyPhred, and PolyBayes and their respective overlaps. The numbers represent the amount of true (TP) and false-positive (FP) SNPs detected by each program, regardless of their quality scores. (*A*) The *SCN1A* data. (*B*) The *MAPT* data.

(Table 1). Even at a quality cutoff of 15, novoSNP detected considerably more SNPs compared to the lowest-quality cutoff for PolyPhred, with a lower false-positive rate than PolyPhred at the highest possible quality (Table 1).

To reject the possibility that the high-quality false positives were, in fact, false negatives from the visual inspection, all high-scoring false positives of novoSNP (quality > 20) and a random sample of 110 highest-scoring PolyPhred false positives (quality = 99) were manually checked, confirming all as true false positives.

Discussion

We developed a software package novoSNP that allows automated detection of sequence variations from sequence trace files in a fast and reliable manner. novoSNP runs on computers with Linux or Windows as operating system. In contrast to assembly-oriented display programs, novoSNP offers a variation-oriented visualization. Important assets of novoSNP over existing variation detection software are: its high rate of correctly identified INDEL polymorphisms, low number of false negatives, availability of an intuitive graphical user interface, and its flexibility in use resulting from the backend database.

Applying novoSNP on a total of 10,090 sequence trace files showed that 511 of the 526 variations identified by visual inspection were detected. This in-house-generated large data set was composed of two independent data sets exemplifying two different approaches: a mutation analysis and a SNP discovery data set. The mutation data set was derived from the exon-based mutation analysis of *SCN1A* in 15 patients within a DNA diagnostic context (Claes et al. 2003). The SNP discovery data set was obtained from the genomic resequencing of 140 kb containing *MAPT* in 23 individuals aiming at generating a high-density SNP map for genetic association studies (Rademakers et al. 2004). Applying novoSNP on the *SCN1A* mutation data set, all 38 variations observed by visual inspection were detected including five INDELs, using a quality score cutoff of 10. Applying novoSNP, with a quality cutoff of 5, on the *MAPT* data set showed that only 15 variations out of 488 (3.1%) were missed, including five INDELs. Using the same cutoff value for both data sets, the percentage of false negatives was always significantly lower for the mutation analysis data set. The other two programs used in this study showed a similar difference in false-negative rate between the data sets (Table 1).

This difference in variation discovery success rate can be explained by the initial scope of the two data sets. Since in DNA diagnosis pathogenic mutations cannot be missed, a rigorous screening is required using more extensive sequence coverage and validated/optimized primer sets, resulting in high-quality sequence trace files. Genomic resequencing, on the other hand, does not necessarily require detection of all SNPs, and such project does not always allow time for optimizing all primer sets for high-quality sequencing and sequence coverage. Indeed, inspection of the unidentified novoSNP variations showed that these were typically positioned near the ends of sequence traces and/or in regions that were quality trimmed by novoSNP.

Performance of PolyPhred and PolyBayes on the same data set showed that even at the lowest-quality cutoffs, both programs missed a large number of SNPs: 175 and 327, respectively (Table 1).

The significantly better performance of novoSNP can be explained by the use of a cumulative scoring scheme that independently examined different variation characteristics. With this approach, variations that have a low score for one characteristic could still be scored if they had a high score for the other characteristics. novoSNP also excelled in the detection of INDELs missing only five out of 41 INDELs in the complete data set. Since PolyBayes does not support INDEL detection, it obviously could not find these. The INDEL detection feature in PolyPhred missed all but two INDELs. Furthermore, novoSNP is not only able to efficiently detect INDELs but also provides the user with the correct sequence of the INDEL.

A high false-positive rate was observed for all three programs used in this study (Table 1; Fig. 2). This is not surprising because the false-positive rate is directly correlated with the overall quality of the sequence traces and especially background noise, and thus is inherent to the discovery methods underlying these software programs. One way to reduce the false-positive rate could be the application of a more consistent selection of PCR and sequencing primers as we did in this study by using the high-throughput primer design program SNPbox (Weckx et al. 2004, 2005). Another way is by relying on the quality scores assigned to the SNP. Indeed, the results presented here showed that the quality score given by novoSNP is a reliable measure of the correctness of the SNP (Table 1). Using a relatively low cutoff score of 10, 97.9% of the true SNPs were found in the combined data sets, but 87.7% of the listed SNPs were false positives. Using higher cutoff values, the number of true variations decreased to 84.3% for a cutoff score of 15, and to 55.7% for a score of 20.

However, the number of false positives decreased accordingly to 32.5% at a cutoff of 15, and only 5.9% for a cutoff of 20. This was not the case for the other two programs, where lower-quality cutoffs detected more true SNPs but the percentage of false positives remained relatively similar with 87.4% to 90.4% for PolyPhred and 78.2% to 84.9% for PolyBayes. This particular feature of novoSNP's quality scores makes it very useful in different scenarios. A high cutoff value can be used when building a SNP map for genetic association studies, where it is important to get reliable SNPs in a prompt manner. Lower cutoff values are to be used in DNA diagnostic mutation analyses resulting in a larger number of marked variations that need followup but at the same time eliminating the risk of false negatives, increasing the probability that pathogenic mutations would not be missed.

To conclude, we showed that novoSNP is an efficient and reliable software package for sequence variation discovery with a high discovery rate of both SNPs and INDELs. Also, the quality score assigned by novoSNP to a marked sequence variant can be used as a reliable criterion for selecting sequence variations for either pathogenic mutation or SNP detection.

Methods

Language and data storage

novoSNP is written in the scripting language Tcl version 8.4, has a graphical user interface written in Tk version 8.4 (http://www.tcl.tk), and can be used on Linux as well as on Windows systems. novoSNP stores all information about reads, alignment, and variations in a single file relational database (http://www.sqlite.org).

Tcl Libraries

novoSNP relies on the additional Tcl packages ClassyTcl, ClassyTk, Extral, and dbi. ClassyTcl is an object system for Tcl and ClassyTk is an extention for Tk, based on the ClassyTcl object system. Additional information can be found at http://www.sourceforge.net/projects/classytcl/. Extral is a library providing additional commands to Tcl (http://www.sourceforge.net/projects/extral/). Dbi is an interface providing a unified way to access different SQL database management systems (http://www.sourceforge.net/projects/tcl-dbi/).

External programs

The BLAST algorithm (Altschul et al. 1990) is used to align sequence reads to a reference sequence. Sequence trace files in ABI format generated on capillary systems like the ABI PRISM 3700 or 3100 DNA Analyzer (Applied Biosystems) or in SCF format generated on other sequencing systems like the CEQ Genetic Analysis System (Beckman Coulter) or the MegaBACE DNA Analysis systems (Amersham) are auto-detected, base-called, and clipped using Phred (Ewing and Green 1998; Ewing et al. 1998). Phred is not included in the distribution of novoSNP but should be obtained according to the information at http://www.phrap.org. Trace files generated on an ABI PRISM 3730 DNA Analyzer require base-calling with the ABI KB BaseCaller. These will be quality-clipped by novoSNP using the same algorithm as used by Phred.

novoSNP strategy

The novoSNP input data consist of a reference sequence in FASTA format, covering the sequenced region(s) as well as the generated sequence trace files. The trace files are preferably arranged per primer set, containing reads from forward and reverse primer sequencing reactions. Once the reference sequence and trace files are added to the single file SQLite database, the program handles all following steps automatically. The trace files are base-called and clipped, and the resulting sequences are aligned to the reference sequence using the BLAST algorithm.

Next, each position in the alignment is scored for the presence of a SNP using a cumulative scoring scheme. The final score for each position is the sum of three subscores, independently determined for forward and reverse reads, and an extra score reflecting how well forward and reverse reads match. The peak size of a color used for the calculation of these subscores is calculated by normalizing the area under this color at the given trace position to the average size of the 16 neighboring trace peaks in the same read (horizontal normalization). The metrics used in calculating the three subscores are illustrated in Figure 3.

The first subscore represents the evidence for a variant in one of the aligned traces ("feature" score). This feature score is calculated by comparing the two largest peaks at the given position for each trace. When only one peak is present (no background), the base represented by this peak will be awarded the maximum score. If two peaks of equal height are detected, both bases will receive a maximum score. Based on a number of cutoffs, lower scores for both bases will be produced when one base

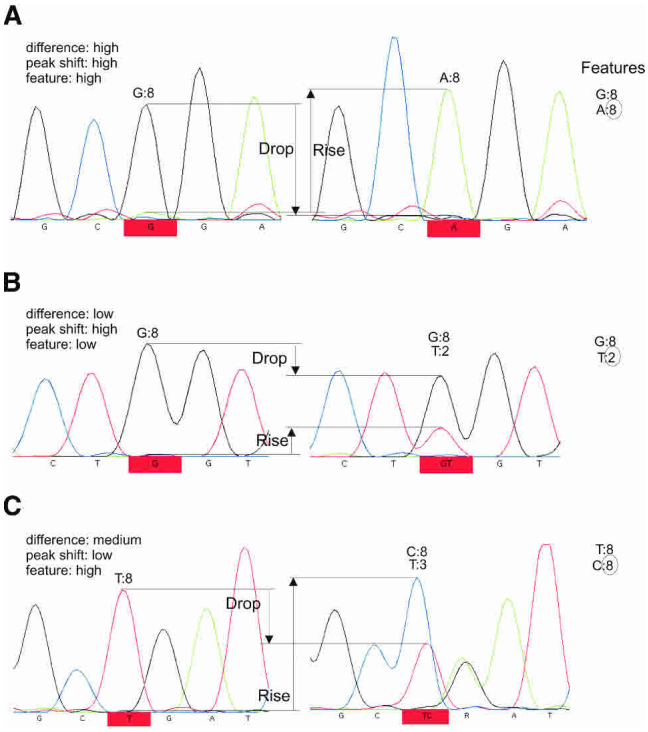

Figure 3. Illustration of the metrics used to calculate the novoSNP quality scores. (*A*) An easy to detect, homozygous SNP that scores high for all metrics. The features seen in the traces and their scores are listed on the *right*; the second best score (circled) represents the feature score for this position. (*B*) An example showing a relative small difference in peak area resulting in a low distance metric. The feature metric assigns a low score to the small secondary peak in the heterozygous sample. However, the compensation in the differences in peak area usually seen in true SNPs and measured by the peak shift metric is very clear: The drop in size in one color is almost just as large as the rise of the other color. (*C*) The peak shift is not always clearly present. In this example, the differences in peak size are large, but their absolute values not similar. The SNP will still be picked up by the other metrics.

peak is at a fraction of the other. When scanning the position in all reads, only the best score for each base is kept, finally resulting in one score for each base at that position. If the position harbors a SNP, two bases instead of one will have high scores. Considering the best scoring base as the wild type, the second best score is the one representing the variant and will be added to the final score as the feature score. Differences with the reference sequence are captured in this score by assigning the reference base a maximum score.

The second parameter is the "difference" score, which provides the largest observed dissimilarity from all combinations of two traces at the given position. The dissimilarity is calculated by summing the peak size differences for each base color. If the highest dissimilarity found exceeds a defined cutoff value, the difference score is added to the final score.

The third parameter or "peak shift" score explicitly targets a typical behavior observed when comparing sequence trace files containing a different allele: A drop in size of a peak in one color is compensated by an equal rise of a peak of another color. The peak shift is calculated by multiplying the ratio between drop and rise of the peaks at the given position with the value of the smallest change at that position. The result is normalized for the highest base peak observed at the aligned position (vertical normalization). If the peak shift between any two reads exceeds a defined cutoff value, a peak shift score will be added to the final score.

In case forward and reverse reads are available, an extra "type" score is added. This extra score reflects how well predicted variants between the forward and reverse reads match. Conflicting reads result in a small penalty. Maximal scores are assigned if the different variants are present in both matching forward and reverse reads.

Heterozygous INDELs show a typical pattern of frameshift-induced stretches containing many double sequence peaks starting at the INDEL position. These regions are clipped before the trace files are aligned to the reference sequence because the base-caller considers these stretches as low-quality bases. Therefore, after SNP scoring of the trace files, each read is tested separately for the presence of heterozygous INDELs by searching for the presence of a double sequence pattern after the clipping point. In case a pattern consistent with the presence of an INDEL is found, the bases corresponding to the reference sequence are removed from the double sequence pattern resulting in a sequence containing the potential INDEL. BLAST alignment of this sequence to the reference sequence determines the presence of an actual INDEL. Heterozygous INDELs are scored according to the length and quality of the aligned part after the INDEL, and the consistency with the double sequence pattern. Finally, homozygous INDELs gathered during the alignment process are evaluated. Since such gaps are often produced by the base-caller missing a call, the quality of homozygous INDELs is determined based on the quality scores of the base-caller, and the uniformity of spacing between the bases in a 6-base region around the INDEL. The uniformity of spacing is scored by comparing the distance between the highest and lowest distance between actual peak positions to the average distance. As with the scoring of SNPs, matching results in forward and reverse reads will add to the final score if available.

The parameters and cutoff values used in the parameter scoring were optimized using large sequencing data sets from resequencing projects in our department.

Sequence trace data sets

We compared the performance of novoSNP with that of PolyPhred (version 4.20) and PolyBayes (version 3.0) for automated sequence variation detection on two data sets, with a total of 10,090 sequence trace files generated on an ABI Prism 3700 and 3730 DNA Analyzer (Applied Biosystems).

The first data set comprises 1028 sequence trace files generated in-house from a diagnostic mutation analysis of the neuronal voltage-gated sodium channel α-subunit type I gene (*SCN1A*), located on Chromosome 2, in 15 patients with severe myoclonic epilepsy of infancy (SMEI). Visual inspection of these sequence trace files identified 38 sequence variations including five INDELs (Claes et al. 2001, 2003; data not shown). Based on the Chromosome 2 genomic sequence with GenBank accession number NT_005403.14 from position 17,050,000 to 17,150,000, we designed primers for this mutation analysis with SNPbox using the "exon" module (Weckx et al. 2005).

The second data set comprises 9062 sequence trace files obtained in-house from genomic resequencing of 140 kb on Chromosome 17q21 spanning the gene coding for the microtubule-associated protein tau (*MAPT*) in 23 individuals to construct a high-density SNP map for genetic association studies (Rademakers et al. 2004). Based on the genomic sequence with GenBank accession number NC_000017.9 from position 41,323,600 to 41,462,800, 198 primer sets were designed with SNPbox using the "saturation" module (Weckx et al. 2005). These primer sets were amplified and sequenced in both directions in 23 individuals. Visual inspection of this data set yielded 488 variations including 36 INDELs. Of 70 variations tested for validation using other technologies, 69 were successfully confirmed.

For novoSNP, a FASTA file of the genomic sequence was used as a reference sequence. For PolyPhred and PolyBayes the reference sequences were translated into a phd file using the fasta2Phd Perl script, and the resulting files were added to the appropriate data sets. PolyBayes was run under standard conditions. PolyPhred was used with default settings except that the option to detect INDELs was enabled. We also ran PolyPhred with a lower-quality clipping score of 15 and with the source option enabled. These settings resulted in a lower number of true variations with a slightly lower false-positives rate. Phrap was used with the force level input variable set to 10 to allow the lowest possible stringency.

Acknowledgments

We acknowledge the contribution of the VIB Genetic Service Facility (http://www.vibgeneticservicefacility.be/) to the resequencing and SNP detection. This work was in part funded by the Special Research Fund of the University of Antwerp, the Fund for Scientific Research Flanders (FWO-V), the Interuniversity Attraction Poles program P5/19 of the Federal Science Policy Office, the Medical Foundation Queen Elisabeth, and the International Alzheimer Research Foundation, Belgium; and an Integrated Project APOPIS within the 6th framework program of the European Commission. R.R. and M.C. are postdoctoral fellows of the FWO-V.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Claes, L., Del Favero, J., Ceulemans, B., Lagae, L., Van Broeckhoven, C., and De Jonghe, P. 2001. De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy. *Am. J. Hum. Genet.* **68:** 1327–1332.
Claes, L., Ceulemans, B., Audenaert, D., Smets, K., Lofgren, A., Del Favero, J., Ala-Mello, S., Basel-Vanagaite, L., Plecko, B., Raskin, S., et al. 2003. De novo SCN1A mutations are a major cause of severe myoclonic epilepsy of infancy. *Hum. Mutat.* **21:** 615–621.
Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.
Fredman, D., Munns, G., Rios, D., Sjoholm, F., Siegfried, M., Lenhard, B., Lehvaslaiho, H., and Brookes, A.J. 2004. HGVbase: A curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res.* **32:** D516–D519.
Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W., and Nickerson, D.A. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23:** 138–144.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23:** 452–456.
Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.
Rademakers, R., van der Zee, J., Bogaerts, V., Van den Bossche, D., Backhovens, H., De Pooter, T., Bel Kacem, S., van Duijn, C., Del-Favero, J., Van Broeckhoven, C., et al. 2004. P4-154 genomic sequencing of MAPT provides an extended SNP map and identifies >30 H1 subhaplotypes. *Neurobiol. Aging* **25 (Suppl 2):** S519.
Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29:** 308–311.
Venter, J., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Weckx, S., De Rijk, P., Van Broeckhoven, C., and Del Favero, J. 2004. SNPbox: Web-based high-throughput primer design from gene to genome. *Nucleic Acids Res.* **32:** W170–W172.
———. 2005. SNPbox, a modular software package for large scale primer design. *Bioinformatics* **21:** 385–387.

Web site references

http://www.phrap.org; Phred/Phrap.
http://www.sourceforge.net/projects/classytcl/; ClassyTcl.
http://www.sourceforge.net/projects/extral; Extral.
http://www.sourceforge.net/projects/tcl-dbi/; Dbi.
http://www.sqlite.org; SQLite.
http://www.tcl.tk; Tcl/Tk.
http://www.vibgeneticservicefacility.be/; VIB Genetic Service Facility.
http://www.molgen.ua.ac.be/bioinfo/novosnp/; novoSNP: a program to find SNPs and small indels in resequencing projects.

Received May 5, 2004; accepted in revised form January 4, 2005.