SEQscoring: a tool to facilitate the interpretation of data generated with next generation sequencing technologies



Katarina Truvé[1], Oscar Eriksson[1], Martin Norling[1], Maria Wilbe[1], Evan Mauceli[2], Kerstin Lindblad-Toh[2,3], Erik Bongcam-Rudloff[1]

[1]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden,
[2]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, USA,
[3]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

Depicted authors have names underlined.

Abstract

Next Generation Sequencing (NGS) technologies promise a revolution in genetic research. Generating enormous amounts of data, they bring both new opportunities and new challenges to researchers. SEQscoring was designed to facilitate analysis and enable extraction of the most essential information from data produced in NGS resequencing projects. Its main functionality is to help researchers locate the most likely causative mutations for a specific trait or disease, but it can advantageously be used whenever the goal is to compare and explore haplotype patterns, and to locate variations positioned in evolutionary conserved genomic elements. SEQscoring uses input data containing information about coverage and variations produced by other programs, like MAQ and SAMtools, and put the emphasis on methods for data visualisation and interpretation. We compare cases and controls in several ways and also utilise the power of comparative genomics, by scoring all variations according to their degree of conservation. SEQscoring is a publicly available, free, web-based service. It has an intuitive interface and can easily be used by biologists, medical researchers, veterinarians as well as bioinformaticians. We exemplify how SEQscoring was used in a recent study as a subsequent step to a genome- wide association study (GWAS) to extract a set of candidate mutations.

Availability: http://www.seqscoring.org

Introduction

"Next generation" sequencing (NGS) technologies are rapidly moving towards faster and cheaper resequencing of whole genomes and transcriptomes [1]. These new sequencing technologies promise to accelerate our knowledge of genetic variation and the associated phenotypic effects. As a consequence we might expect disease-causing mutations to be revealed and to see an advance in therapies and development of individually tailored drugs [2]. To be able to interpret the vast amounts of data being generated, new tools and algorithms will be needed for extensive comparison of entire individual genomes.

Resequencing not of entire genomes but of targeted regions has quickly become a valuable strategy to find candidate mutations following identification of associated regions using genome-wide association studies (GWAS). When performing GWAS, the use of SNP-chips, with thousands or several hundred thousands of single nucleotide polymorphisms (SNPs) evenly spread over the genome, makes it possible to locate disease-associated regions. Locations where allele frequencies differ between "cases and controls", may indicate a region harbouring a mutation where cases are identical by descent. Usually, a denser fine mapping of the located region(s) follows the GWAS. These methods have proven successful for identifying mutations inherited in a Mendelian fashion [3]. Most disease-causing mutations have been found in exons, probably because they have been subject to the most intense investigation, their causative effects being easier to validate than those of mutations in other regions. Yet, many regions outside exons have important regulatory effects, for example with respect to the location, timing and amount of gene expression. Particularly in complex diseases, where several genes and also environmental factors are involved, regulatory mutations are likely to be common. NGS allows the detection of variants in a wholly new scale. Consequently, we expect important mutations to be revealed with higher frequencies using these new methods, not just for those located in exons.

With new opportunities also come new challenges. The large amount of variation present in every individual (~1/1000 bp) raises the question of how to 'separate the wheat from the chaff'. The approach we outline here makes use of comparative genomics, as it has been shown

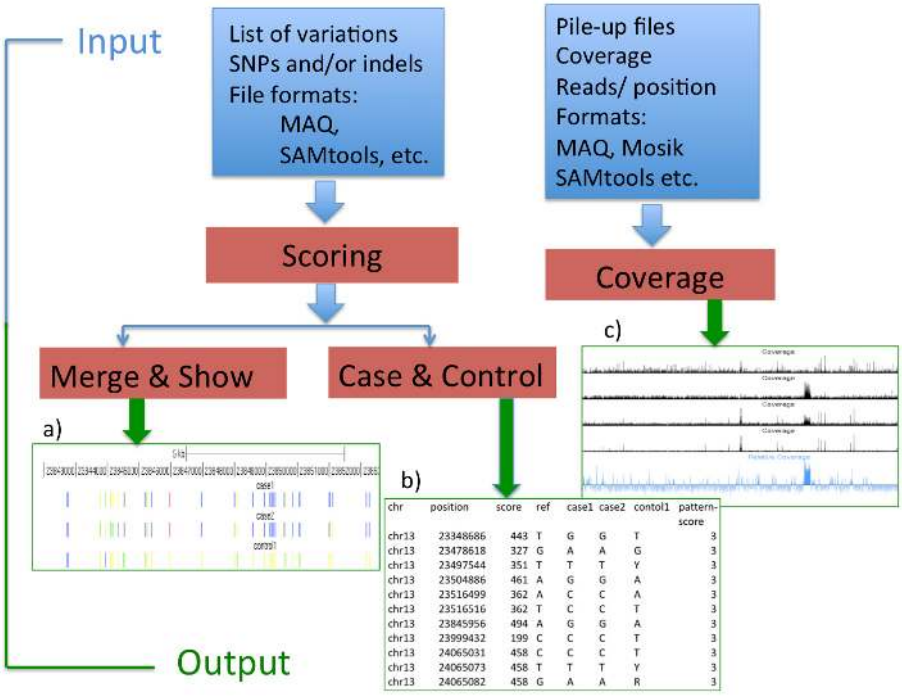

Figure 1. Overview of SEQscoring modules
The user submits input data in the form of lists of variants or coverage data produced by programs like e.g. MAQ or SAMtools. Variation data (SNPs/indels) are first scored by the Scoring module according to the degree of evolutionary conservation at the genomic position for the variation. In the next step data can be visualised using the Merge & Show module that aims to facilitate study of haplotype type structure and conservation within haplotypes. The Case & Control module, takes sample phenotype into account and aims to find the most likely positions or regions harbouring a causative mutation. The Coverage module performs calculations to find differences between cases and controls in an attempt to localize structural variants, like deletions or duplications. The output is provided in lists for download and further analyses and possibilities for direct visualisation in the UCSC browser. Some examples are shown in the figure; a) colour coded SNPs with a similar haplotype in cases, one SNP (red) within a conserved element; b) a table of conserved SNPs with calculated pattern-scores; c) visualisation of coverage differences between cases and controls.

that elements that are conserved across species, and are thus under purifying selection, are more likely to have a function [4-7]. We have therefore developed a tool, SEQscoring that scores mutations according to the degree of conservation, and also takes the pattern between cases and controls into account. The tool is freely available and can be accessed via the Web. The program also aims to facilitate the identification of structural variations, such as deletions and duplications, by calculating the ratio of average coverage between cases and controls in windows of a specified size. To allow comparison of individual datasets, some results are provided in a format compatible with the UCSC genome browser [8]. To facilitate the interpretation of data overall, SNPs and indels (small insertions or deletions) are colour coded in such a way as to give users an overview of features like homozygosity, conservation and variation. SEQscoring has been tested on several data sets, and shows great potential to help the user to extract the most essential information from their NGS-projects. The tool is easy to use and has an intuitive interface that can be used by biologists, medical researchers and bioinformaticians.

Results

Design and implementation

The SEQscoring tool aims to study haplotype structure and to localise important differences between cases and controls in genomic regions where NGS data are available. In Figure 1 we give an overview of the SEQscoring modules. The modules are described in more details below.

Prior to SEQscoring, variant detection should be performed using state of the art methods. Several different programs can be used to map millions of read to a reference sequence and to call variations, e.g. MAQ [9] and SAMtools [10]. SEQscoring supports several different file formats as input data and our ambition is to include additional formats if requested. Typically we expect

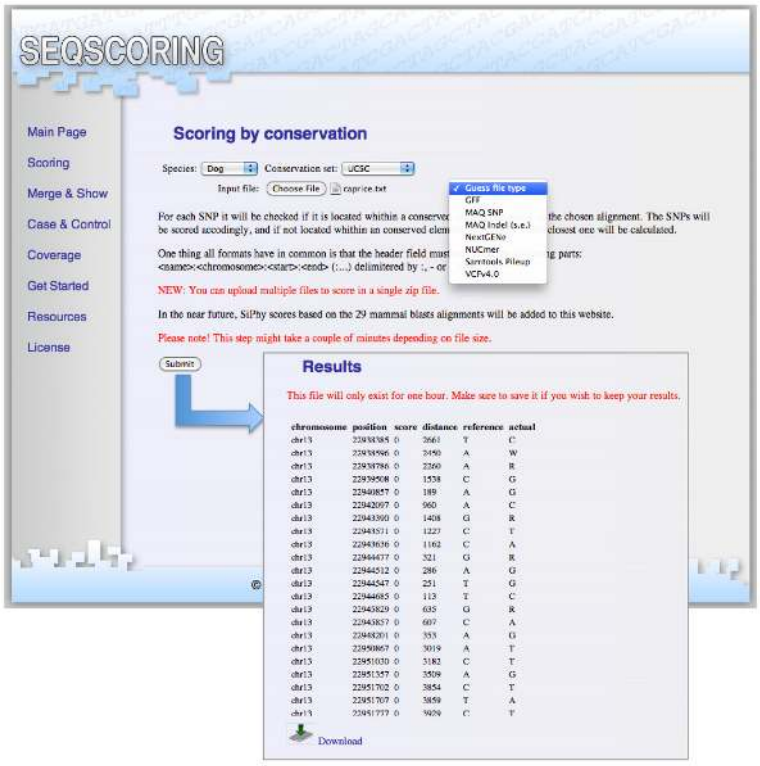

Figure 2. Conservation scoring identifies the variants with constraints and therefore with a higher chance to have a phenotypic effect. At the scoring page the users can submit their files of variations after choice of species and alignment/method for finding conserved elements. In the output file each variation has got a conservation score, and if not within a conserved element the distance to the closest one has been calculated.

the user to submit a single list with variants (SNPs and/or indels) for each individual.

To make this service accessible, it has been implemented as a web site hosted by an Apache web server running Python and Perl CGI scripts. Due to the dynamic content presented on the pages, the scripting language PHP is utilised for creating web pages. Python CGI scripts are used to catch both the raw data and the parameters from each form. Perl or Python modules then carry out the data processing. All data uploaded and produced by SEQscoring is stored for a limited period of time and then automatically erased. Submitted files get unique encrypted file names using MD5 sums in order to minimize the risk of access by unauthorized users.

Conservation scoring

In the Scoring module, variants are scored according to the degree of constraint at the genomic location for the variation. In principle, data from any species can be analysed as long as constraint score data is available for the particular species. For each variant the scoring module checks whether it is located within a constraint element, and if not the distance to the closest one is calculated. The location of constraint elements may differ depending on method and species used in the alignment. For mammals we propose the use of the 29mammals constraint scores (SiPhy omega or pi [11-12]) lifted onto the respective genome. Other available datasets are 16 amniota vertebrates and human/mouse/rat/dog comparison (Pecan [13] and PhastCons [14]). Those records are kept in our local database for high performance. Python modules performing iterative binary search have been implemented and compiled as C-extensions for fast and memory efficient conservation scoring of user submitted variations.

Visualization of variation and conservation

The Merge & Show module merges all variants and their score for all individual into a text file that can be downloaded for further analyses. The data is also displayed in the UCSC genome browser for easy comparison and investigation of haplotype structure. For easier interpretation SNPs

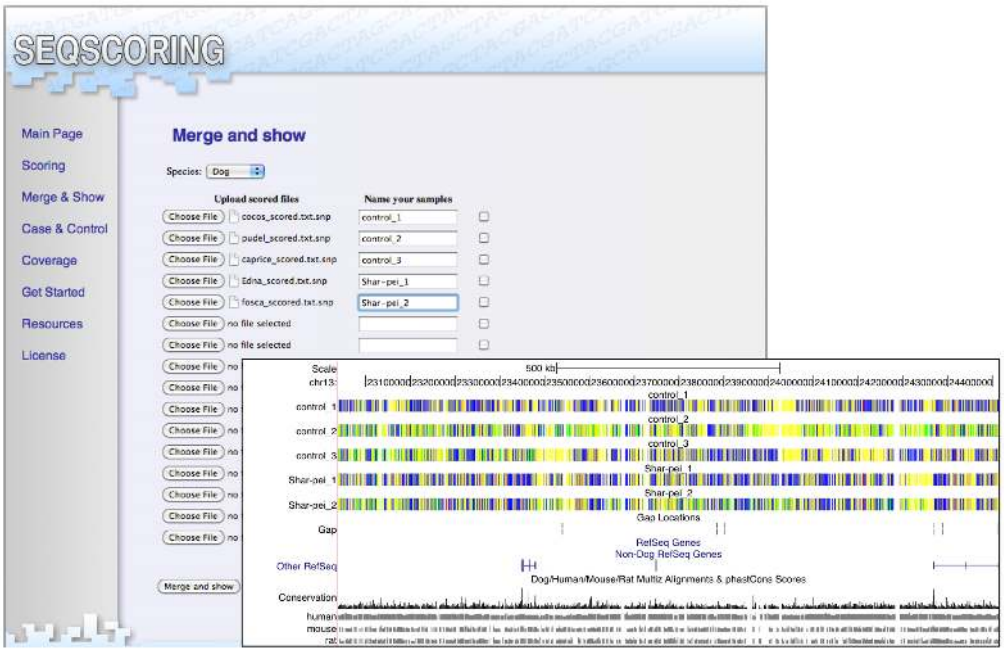

Figure 3. Merge samples and show variations in the UCSC Genome browser.
To ease the comparison of samples there is an option to merge and display scored files in the UCSC Genome Browser. Here an example of SNPs from five dog samples is displayed in the browser. SNPs are colour coded in the following way: yellow for homozygous equal to reference, blue for homozygous deviating from reference, green for heterozygous, red for homozygous within (± 5bp) a constraint element and pink for heterozygous within (± 5bp) a constraint element.

are displayed with the following colour code: homozygous SNPs within or near (± 5bp) constraint elements are coloured red; heterozygous SNPs within or near (± 5bp) constraint elements are coloured pink; non-constraint homozygous SNPs corresponding to the reference allele are coloured yellow; homozygous SNPs deviating from the reference are coloured blue; heterozygous non-constraint SNPs are coloured green.

Evaluation of concordance with phenotype status

The Case & Control module gives further help to reveal differences between cases and controls. Three different options are offered the user: 1) to compare constraint variants; 2) to compare genomic regions; 3) to transform data into a format for doing traditional association studies.

The first option selects SNPs located in conserved elements and scores them according to concordance with an expected pattern. The algorithm goes through all possible combinations of individual pairs and calculates a pattern-score depending on what the expected pattern of alleles are taking mode of inheritance into account. Cases and controls can be defined either based on phenotype or genotype expectation. We consider the highest scoring variants identified in this way to be among the most likely to be causative of the trait under investigation. Pattern-scores for conserved SNPs are calculated in the following way:

n = set of all samples
i= genotype for sample 1
j =genotype for sample 2
S(i) = status for sample (case or control)
p= pattern-score

$$p(SNP, I) = \left| \left\{ \begin{array}{c} (i,j) : i,j \in n \wedge \\ S(i) \neq S(j) \wedge i \neq j \vee \\ I = recessive \wedge S(i) = case \wedge S(j) = case \wedge i = j \vee \\ I = dominant \wedge S(i) = control \vee S(j) = control \wedge i = j \end{array} \right\} \right|$$

The second option "compare genomic region", scans for regions of specified size where cases are alike and differ from controls. Pair-wise combinations are examined in a similar way as for conserved SNPs, but all SNPs conserved and not conserved are taken into consideration. The mode of inheritance is not taken into consideration here. A sliding window approach is used and the highest score goes to the region that is as homozygous as possible in cases, and differ as much as possible to the controls. This option can be used to look for selective sweeps as well as for

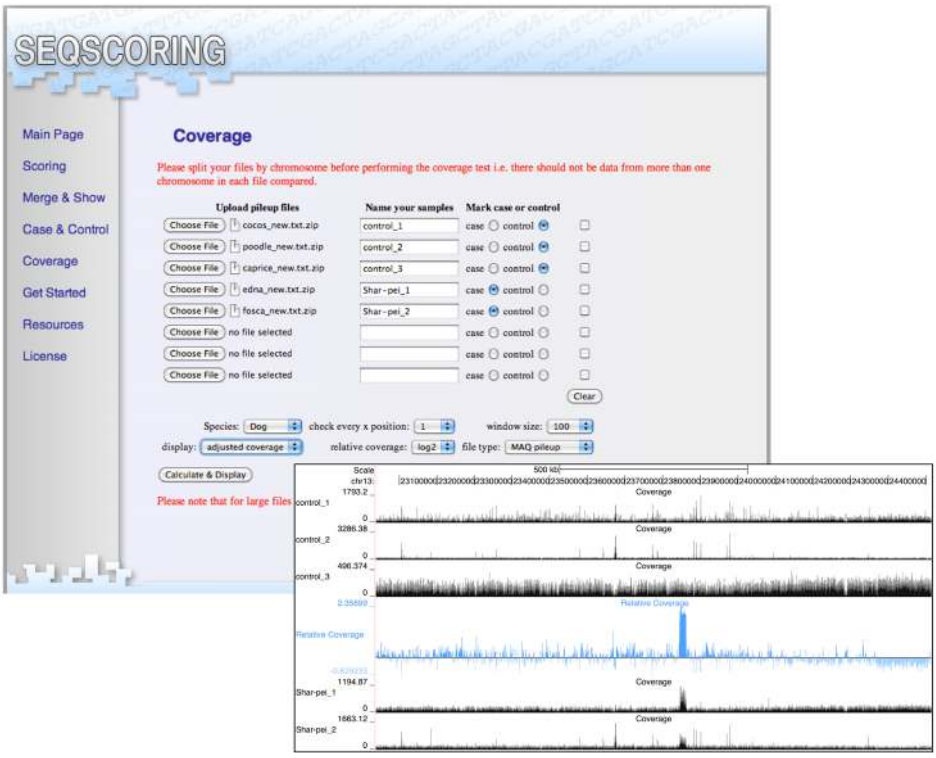

Figure 4. Comparison of coverage may reveal copy number variations.
The coverage option can be used to visualise differences in average coverage between cases and controls in an attempt to localize structural variations as duplications or deletions that might be causative for a certain phenotype. Here we show five samples from different dog breeds. Two of them are Shar-pei dogs with a thick wrinkled skin phenotype. The blue graph shows the coverage ratio (log2) between Shar-pei and the control breeds. Shar-peis clearly show a peak of excessive coverage that has now been proven to be a 16 kb duplication affecting both the skin phenotype and a fever disease in Shar-pei dogs.

smaller homozygous regions that might be identical by descent in cases harbouring a possible causative mutation for a specific trait.

Relative coverage analysis

The Coverage module is aimed to identify structural variation and at present it accepts pileup files created by MAQ [9], SAMtool [10] or Mosaik [15]. If there is a big difference in average coverage between data from different samples, the data can be normalised by setting the average coverage to the same fixed value for each individual. Comparable figures are thus calculated by dividing all data with an individual adjustment factor. There is also an option to average the coverage in a window of a specific size. The ratio of coverage between cases and controls is calculated for windows of a specified size and log2 transformed. The number of positions checked is limited to 150kb due to performance, thus giving a maximum resolution to regions smaller or equal to that size and subsequently diminishing resolution for larger regions.

Example

SEQscoring is currently in use at our lab for several resequencing projects where the aim is to find mutations responsible for specific traits or diseases in dogs. In our group, we traditionally use the dog as a disease model. The results obtained may, in many cases, be successfully translated to humans, and the knowledge gained thus has the potential to benefit both species [16-19]. It should be noted, however, that the methodology and software tool that we present here are generic, and not tied to a specific species.

In one of the first NGS projects at our lab the aim was to find the mutation responsible for the characteristic wrinkled skin phenotype in the Chinese Shar-pei dogs, a phenotype strongly selected for by breeders. The breed also suffers from a genetic disorder called Familial Shar-pei fever, a disease resembling human hereditary periodic fever syndromes. It has now been shown that the two features are connected and caused by a pleiotropic mutation [20]. We here exemplify the use of some SEQscoring functions with data

from the Shar-pei project. A region of 1.5 Mb had been selected for resequencing based on a genome wide SNP analyses showing strongly reduced heterozygosity in Shar-peis, implicating the presence of a selective sweep. The region was captured using custom designed arrays from NimbleGen and sequenced using Illumina Genome Analyzer. The obtained sequence reads were aligned to the target region of CanFam2.0 [21] using MAQ [9]. In Figure 2 it is illustrated how called SNPs are scored by conservation using SEQscoring. In this example we chose the UCSC phastCons alignment of four species. In the output file each variant has got a conservation score and, if not within a conserved element, the distance to the closest one has been calculated. In the first sequencing experiment two Shar-peis and three control breeds were sequenced. When the reads were mapped to a repeat masked reference ~1500 SNPs/individual were detected. In the next step we used the Merge & Show module. Downloading a text file with all SNPs/individual merged let us count that there were 3,430 SNPs in total and out of those only 84 were within conserved elements. The results are displayed in the UCSC genome browser (Figure 3) with colour coding as explained above. Next we used the Case & Control option to compare conserved SNPs, and found that only eight of the conserved SNPs had a pattern where the two Shar-peis were alike and differed from the controls. Those eight SNPs have been genotyped in several samples and in this case shown not to be causative since they were not unique for Shar-peis. Finally, we use the Coverage module to explore if there are any coverage differences between Shar-peis and controls. We targeted a region of 1.5 Mb and maximum of 150kb are displayed, meaning that in this case the program check the coverage at every 10th position. We also chose to use adjusted coverage and to average the coverage in a window size 100, actually meaning 100 *10 (every 10th position checked) = 1000 bases window. Coverage graphs were directly displayed in the UCSC genome browser. As can be seen (Figure 4) there was one clear peak of excessive coverage in both Shar-peis. The blue graph shows the log2 values of the ratio between cases and controls. It has now been shown that Shar-peis have a 16.1 kb duplication at this site [20].

Discussion

We have demonstrated how the use of the publicly accessible SEQscoring web site facilitates the interpretation of data from NGS-projects. We expect that the user, in most cases, is interested in localising the mutation for a specific phenotype. For best use of resources we propose a model where a number of individuals (6-12) are picked for resequencing, consisting of both cases and control. The region suspected to harbour the mutation has been narrowed down by GWAS before NGS.

It is assumed that genomic regions that are conserved across species are under evolutionary constraint and thus more likely to be functional. For this reason SEQscoring offers a fast conservation filtering of user submitted variations. It is important to be aware that different algorithms, and the set of species represented in the alignment, are likely to find different constraint elements. At present two different sets of constraint elements can be used for filtering. We are planning to add a third set in the near future, where conserved elements have been identified by alignment of 29 mammals using SiPhy [11]. In addition, a candidate function have been suggested for up to 60% of constrained bases [12]. We think that conservation filtering is an important and valuable step in variation evaluation but it should also be kept in mind that sometimes, functional elements show low degree of sequence conservation.

As mentioned in the results section there is an option to transform the NGS data to a format that can be used for traditional association studies based on allele frequencies using the program PLINK [22]. Usually a small number of samples are under investigation by resequencing, and the sample size is not appropriate for large-scale association. However, in the case that a larger sample size is utilised we offer a down load format that allows export of data into plink format. We offer two other methods to evaluate the concordance of genotype with phenotype: to compare conserved SNPs, and to compare genomic regions that have been designed with the purpose to extract as much information as possible using relatively few samples. The option to compare conserved SNPs uses both the power of conservation filtering and the identification of a pattern in concordance with an expected mode of inheritance thus capable of extracting the most likely causative SNPs for a specific trait.

The option to compare genomic region would most likely find homozygous regions containing two risk alleles (homozygosity), and is therefore most applicable to recessive traits and traits under selection. Dominant traits and complex risk factors are harder to identify. Usually cases and controls are defined based on phenotype, but as haplotype information from the GWAS or fine-mapping is typically used when picking samples for resequencing, to increase the odds of localising the causative mutation we recommend to use controls that are believed to be homozygous for an assumed healthy wild-type haplotype.

Sometimes structural aberrations like insertions, deletions or duplications are responsible for a specific trait. The possibilities to detect such differences are limited in resequencing projects. If there is an insertion in one of the individuals, those reads will simply not mapped to the reference. The read-length is often quite short (~30-100 bp) limiting the size of repetitive regions that can be read through, meaning that differences in size of microsatellites, presence of LINES and SINES etc.,can be hard to detect. We propose the use of tools that do de novo assembly to be able to capture putative insertions and deletions. After assembling larger contigs those could be mapped back to the reference and thus detecting insertions, but still the maximum detectable insertion size would be approximately the size of the read length.

The use of paired end reads offers a possibility to detect larger insertions, duplications and deletions but will not recognise smaller differences since those might be due to different shearing size. For single end reads, the information about coverage at each position has proven to be useful for identifying the putative locations of copy number variation or deletions that differs between cases and controls.

NGS projects are likely to identify a vast amount of variations between individuals and it is challenge to extract the ones that might be functional. We propose a methodology where the goal for the analyses is to extract a limited set of variations most likely to be causative for the trait under investigation, and to continue the analyses by genotyping in several cases and controls. We showed that the analyses offered by SEQscoring are straightforward and easy to understand, but powerful and time saving through the ability to extract important information, visualise the results and help the user propose a set of candidate mutations from the vast amount of data produced.

Acknowledgments

We thank the 29mammals consortium for allowing us access to the SiPhy mammalian constraint elements. This work has been supported by the EMBRACE project funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092 and Bioinformatics Infrastructure for Life Sciences (BILS) funded from the Swedish Research Council. KLT is funded by a EURYI from the ESF.

Competing interest statement

None declared

References

1. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nature biotechnology 26: 1135-1145.
2. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nature genetics 39: 1522-1527.
3. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322: 881-888.
4. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS biology 3: e7.
5. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nature genetics 38: 223-227.
6. Margulies EH, Blanchette M, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome research 13: 2507-2518.
7. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
8. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome

Browser database: update 2011. Nucleic acids research 39: D876-882.

9. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research 18: 1851-1858.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
11. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, et al. (2009) Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25: i54-62.
12. Lindblad-Toh K GM, Zuk O, Lin M.F , Parker B.J (2011) A high-resolution map of evolutionary constraint in the human genome based on 29 eutherian mammals. Submitted.
13. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome research 18: 1814-1828.
14. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research 15: 1034-1050.
15. MOSAIK the reference-guided assembler: [http://bioinformatics.bc.edu/marthlab/Mosaik]
16. Karlsson EK, Lindblad-Toh K (2008) Leader of the pack: gene mapping in dogs and other model organisms. Nature reviews Genetics 9: 713-725.
17. Patterson DF, Pexieder T, Schnarr WR, Navratil T, Alaili R (1993) A single major-gene defect underlying cardiac conotruncal malformations interferes with myocardial growth during embryonic development: studies in the CTD line of keeshond dogs. American journal of human genetics 52: 388-397.
18. Mellersh CS, Boursnell ME, Pettitt L, Ryder EJ, Holmes NG, et al. (2006) Canine RPGRIP1 mutation establishes cone-rod dystrophy in miniature longhaired dachshunds as a homologue of human Leber congenital amaurosis. Genomics 88: 293-301.
19. Green SL, Tolwani RJ, Varma S, Quignon P, Galibert F, et al. (2002) Structure, chromosomal location, and analysis of the canine Cu/Zn superoxide dismutase (SOD1) gene. The Journal of heredity 93: 119-124.
20. Olsson M, Meadows JRS, Truvé K, Rosengren-Pielberg G, Puppo F, Mauceli E. (2011) A Novel Unstable Duplication upstreams of HAS2 predisposes to a Breed-defining skin Phenotype and a Periodic Fever Syndrome in Chinese Shar-Pei Dogs. PLoS Genet 7(3):e1001332.
21. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438: 803-819.
22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81: 559-575.