Genetics and population analysis

Computational identification of candidate loci for recessively inherited mutation using high-throughput SNP arrays

Marko Laakso¹, ³, Sari Tuupanen², ³, Auli Karhu², ³, Rainer Lehtonen², ³, Lauri A. Aaltonen², ³ and Sampsa Hautaniemi¹, ³, *

¹Computational Systems Biology Laboratory, Institute of Biomedicine, ²Department of Medical Genetics, and ³Genome-Scale Biology Research Program, Biomedicum Helsinki, University of Helsinki

Received on November 3, 2006; revised on April 22, 2007; accepted on May 10, 2007
Advance Access publication May 17, 2007
Associate Editor: Keith Crandall

ABSTRACT

Motivation: Single nucleic polymorphisms (SNPs) are one of the most abundant genetic variations in the human genome. Recently, several platforms for high-throughput SNP analysis have become available, capable of measuring thousands of SNPs across the genome. Tools for analysing and visualizing these large genetic data sets in biologically relevant manner are rare. This hinders effective use of the SNP-array data in research on complex diseases, such as cancer.

Results: We describe a computational framework to analyse and visualize SNP-array data, and link the results in relevant databases. Our major objective is to develop methods for identifying DNA regions that likely harbour recessive mutations. Thus, the algorithms are designed to have high sensitivity and the identified regions are ranked using a scoring algorithm. We have also developed annotation tools that automatically query gene IDs, exon counts, microarray probe IDs, etc. In our case study, we apply the methods for identifying candidate regions for recessively inherited colorectal cancer predisposition and suggest directions for wet-lab experiments.

Availability: R-package implementation is available at http://www.ltdk.helsinki.fi/sysbio/csb/downloads/CohortComparator/
Contact: sampsa.hautaniemi@helsinki.fi
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Inherited genetic variation plays a key role in several diseases, such as cancer. With the advent of human sequence project (International Human Genome Sequencing Consortium, 2004) it is possible to delve into genetics and unravel associations between genetic variations and diseases. An important research direction is to compare cases to controls and identify genes for which both copies are missing or mutated as these genes are candidates to cause recessive predisposition to the disease. For example, in cancers tumour suppressor and caretaker genes encode proteins that regulate the growth and genomic stability of the cell. If both alleles of a tumour suppressor gene are mutated and become non-functional, somatically or in germline, this may accelerate cancer genesis.

When trying to identify genes whose malfunction leads to accelerated disease progression, the concept of homozygosity becomes very important. For example, phenotypic changes caused by tumour suppressor gene defects often require homozygosity in the corresponding DNA region because the inactivation of a gene in one chromosome can be usually compensated by the unaffected gene in the other chromosome. Furthermore, homozygosity can be used to measure linkage disequilibrium and locate disease gene candidates (Sabatti and Risch, 2002).

Ideally, homozygote regions are identified using the whole DNA sequence, but to date this is prohibitively costly. Thus, the current practice is to use markers that are distributed across the DNA sequence. One of the most high-resolution marker approaches to date is based on single nucleotide polymorphisms (SNPs); DNA sequence variation occurring in a single nucleotide. Single nucleic polymorphisms (SNPs) are one of the most abundant genetic variations in the human genome (Mooney, 2005). Hundreds of thousands of SNPs across the whole genome can be queried with high density SNP-microarrays.

Currently, there are only few systematic tools for SNP-microarray data analysis, for example, dChipSNP (Lin, et al., 2004) and SNPscan (Ting, et al. 2006). These methods focus mainly on statistical analysis of loss-of-heterozygosity (LOH) and chromosomal copy number aberrations, which are different problems from identifying homozygote regions. In the context of identifying weak signals, a popular option to analyse genetic data, hidden Markov models (HMMs), are not very attractive because they tend to produce ambiguous bounding for the homozygous regions depending on the iteration direction along the SNP sequence. The HMMs also require training data that may not be easy to obtain to estimate several parameters, which sometimes do not have clear correspondence outside the HMM.

In this study, we describe a two-tier approach to mine relevant disease associating genes from high-throughput SNP data. First, the methods to identify biologically interesting
regions are implemented as a module (CohortComparator) that can be used for the rapid and integrated analysis of SNP-microarray data. Second, we have constructed a framework (RegionAnnotator) for annotating the genes from CohortComparator. Together these modules offer means for rapid and integrated analysis of SNP-microarray data. Further, we have developed a novel way to visualize the SNP-microarray data, which differs from dChipSNP in a sense that our emphasis is on the alleles of the homozygous regions and their relation to the corresponding alleles in references. CohortComparator detects homozygous regions similar to the so called NumHom method (Beroukhim, et al., 2006), but unlike the NumHom method, CohortComparator takes the genotyping errors into account. Moreover, the methods in Beroukhim et al. (2006) are designed to down-weight linkage disequilibrium regions, while here we are primarily interested in these regions.

Our main objective is to transform high-throughput multivariate genetic data to biologically testable hypotheses. CohortComparator enables systematic analysis and visualization of the genetic high-throughput data sets in order to achieve this goal. Accordingly, we have implemented several computational methods to identify homozygous, deleted and compound heterozygous regions. Furthermore, we have developed tools to rapidly annotate the results via relevant biological databases, such as Ensembl (Birney, 2006), Gene Ontology (Ashburner, 2000), and SNPs3D (Lin, 2004). We provide a proof-of-principle study using SNP-microarray data from colorectal cancer patients and controls. The methods are implemented with Java and R statistical software (R Development Core Team, 2006) and are freely downloadable at our website.

2 APPROACH

We suggest a framework to analyse genotype data and to identify a set of candidate genes to be associated with a recessively inherited phenotype. An overview of the procedure is illustrated in Figure 1. The need for the framework we describe in this study stems from two sources. First, as the number of SNPs to be genotyped in a very large patient cohort is limited, there is a need for tools that suggest the most likely candidate SNPs for these large-scale studies. Second, statistical tests may not be able to find weak signals from relatively small data sets (Houlston and Peto, 2004). For example, in cancer research this would mean to identify homozygous regions shared by only 2–3 patients but which are missing from all reference samples using few hundred samples. That is, a very sensitive methodology with relaxed requirements for specificity is needed. The lack of specificity is overcome by fusing information from biological databases. The result of the analysis is a set of regions that are mostly likely harbouring biomedically relevant transcripts.

In general, when identifying recessive gene variants the use of homogenous populations like those in Finland, Iceland, Quebec and Northern Sweden is very useful and sometimes the only functional strategy to detect rare genetic variations. In a mixed population the background homozygosity rate is lower, and therefore a signal caused by recessively inherited predisposition genes can be detected more easily in homogenous populations. Thus, the implemented methods are especially suitable for isolated populations descending from small populations but are applicable to more heterogeneous populations as well. Further, the methodology we develop in this study is most suitable if the assumption that genotype samples originate from distantly related individuals who
express the phenotype in question and from individuals who do not express the phenotype is invalid. A schematic of the procedure is illustrated in Figure 2. The actual comparison of the data sets may involve iteration of the parameters until the results are ready for integration. The analysis results in a list of homozygous regions in patient samples that are not present in the reference samples. This list can be used as an input for the downstream applications that can couple them with gene annotations, pathway information and so forth.

Homozygous regions of an individual’s genome can be found from the SNP genotypes. An SNP genotype data set consists of sample vectors of $k$ diallelic SNP measurements. Each measurement is represented as a realization from the set $\lambda = \{AA, AB, BB, NoCall\}$, which corresponds to the detected alleles of both homologous chromosomes and missing values. In order to enable data integration across several high-throughput SNP platforms, we have included a haplotype estimation procedure for the situations where SNP density is different between the two groups to be compared.

2.1 Cohort comparison

The definition for a homozygous or biologically interesting region is dependent on the purpose of the study and data. Therefore, we have implemented the algorithms so that the parameters used can be chosen according to the goals of the study. However, practically always a continuous sequence of homozygous SNPs is considered as a homozygous region, and the question is to decide the minimum length of consecutive homozygous SNPs ($l$), how large ($s$) and dense genotyping errors of heterozygous SNPs are tolerated before declaring a region heterozygous. The allowed density of heterozygous gaps—tolerated because calling errors are bound to occur in the high-throughput data—is determined by a sliding window that calculates the rate of heterozygous SNPs within the $w$ closest SNPs. A SNP is considered to be a part of a homozygous region if it is homozygous or the gap density is below the given threshold $g$ ($0 \leq g < 1$).

The homozygous and heterozygous regions can be represented with a Boolean matrix that contains ‘1’ if the corresponding SNP is within a homozygous region and ‘0’ otherwise. Given $n$ samples and $k$ SNPs in a chromosome, we define the matrix $R_{sw} \in \{0, 1\}^{n \times k}$ of homozygous regions. In order to compute $R_{sw}$, we first need to calculate a matrix in which the entries ‘1’ and ‘0’ correspond to homozygous and heterozygous SNPs, respectively ($R \in \{0,1\}^{n \times k}$). The matrix $R_{sw}$ is constructed using the sliding window to fill gaps that are smaller than $s$:

$$R_{sw} = \left( \frac{\min_{1 \leq c \leq w} \left( \frac{1}{2} - 1, k \right)}{w} \right) \geq 1 - g \quad R_{sw, w} = \begin{cases} 1 \quad & \text{if } w \in [1, k] \end{cases}$$

Second, the continuous regions are restricted to those with at least $l$ SNPs:

$$R_{sw} = \left\lfloor \frac{c}{l} \right\rfloor \left( \sum_{i = \max(1, l)}^{\min(c + l - 1, k)} R_{sw, i} \right) = l, \quad l \in [1, k].$$

Homozygous regions are divided into shorter regions based on their allele constitutions. The distinction between the most common and rare alleles facilitates the detection of a possibly mutant carrying allele. Vector $\gamma$ contains all SNPs that belong to at least one homozygous region found from the reference samples. Each SNP in $\gamma$ is represented by AA or BB depending on which genotype is the most common within the homozygous regions in the reference set. Heterozygous gaps and NoCall SNPs are ignored. Genotype AA is assumed if the frequency of both genotypes is equal. The splitting of the homozygous
regions is based on the comparison between the sample genotype and $\gamma$. The regions between SNPs matching $\gamma$ are labelled as common alleles and other regions represent rare alleles. The base-pair division between the haplotype boundaries favours the rare regions so that a rare allele is expanded to the limiting SNPs. By assuming that the rare genotype is missing in the reference set we try to maximize the sensitivity of detecting allele differences.

Biologically interesting homozygous regions can be found by excluding the regions that are homozygous in reference samples ($R^R$) and by requiring an overlap of multiple homozygous regions in patient samples ($R^P$). Homozygous regions of the references are compared against $R^P$ to ensure that the alleles of the regions are the same. A region is labelled as interesting if the rare allele is absent in the reference samples but present in multiple copies in $R^P$. The rate of required copies can be given as a ratio of samples in $R^P$ and a rate of tolerated findings can be given for the samples $R^R$.

A separate comparison of NoCall-values follows the same pattern as that of the homozygous regions. The sequences of at least $d$ NoCall SNPs are recognized as possible homozygous deletions. The deletion sequences of both cohorts are compared in order to detect those deletions that are frequent in $S^D$ but not present in $S^R$.

Regions of compound heterozygosity contain two different mutant alleles at the same locus. Individuals with compound heterozygote regions may express recessive phenotypes if the both alleles of the putative predisposition gene are deficient. The interesting homozygous regions of $R^D$ can be compared against the other samples to detect possible compound heterozygotes. The further analysis of these compounds may reveal interesting alleles that are not associated with their compound in references.

The results of the cohort comparison are reported in two ways. First, the exact base pairs of each homozygous and interesting region with the associated sample names are written into text files that can be used to track down the original SNP measurements. Other programs, such as the annotation module implemented here, can use these files to query additional information from biological databases. Second, the findings are visualized in graphs such as Figure 3. The graphs represent homozygous regions and their alleles. Each sample has its own row, whereas the X-axis represents the base pairs. The reference cohort is shown on the top of the image followed by the patient samples that are sorted based on the total length of the heterozygous regions. Candidate regions for homozygous deletions are illustrated with black bars. The physical location and the type of the interesting regions can be seen from the

Fig. 3. CRC patient samples compared against the reference samples. Allele colours of the homozygous regions have been chosen so that red represents the $\gamma$ allele and the other allele is blue. Dark violet is used as a compromise in those experimental regions that have no colour determining SNPs, which means that they are on a region without homozygous regions in reference set. Two dark violet regions may differ in their allelic constitution independently. Regions with possibly interesting differences have been emphasized with vertical blocks covering the regions. Homozygosity has been marked with yellow, deletions with green and exceptional alleles with orange. The possible compound heterozygotes have been shown in grey as overlapping interesting regions. Patient samples have been sorted based on the total length of their homozygous regions and the distribution of the total lengths is compared against the normal distribution using Shapiro–Wilk normality test (Shapiro and Wilk, 1965).
2.2 Scoring
The methods for detecting biologically interesting regions are designed to have as high sensitivity as possible in order not to miss potentially repressed genes. The drawback of this design is a large number of false positives. In order to identify those regions that most likely harbour true positives, we have developed several ways to exclude the false positives. In this section, we describe a scoring method that ranks the homozygous regions according to their characteristics.

A possibly interesting region consists of samples with homozygous alleles overlapping the region loci. The relevance of the region depends on its length and relative frequency of its alleles between the controls and the patient samples. Using these measures we have developed a method to suggest interesting region candidates and rank all regions according to the likelihood that they indeed are interesting. A homozygous region with an allele not homozygously present in the reference data may cover the whole interesting region length in one sample, but another sample may share a fraction of the homozygous allele. Scoring of a region is based on the length of the features and thus we are not calculating the actual length of the region itself but the total length of overlapping sample features in it. The actual sum of the overlapping features is divided by the cohort size to overcome potential bias introduced by the cohort size. An example of the scoring approach is shown in Figure 4 that consists of two interesting regions A1 and A2. In this example, the interesting features are required from two or more samples and they must not be present in the reference. A1 consists of alleles with black background present in homozygous regions r3, r4 and r8. A2 represents the allele-unspecific homozygosity of homozygous regions r5, r7 and r9. Lengths associated with scoring, are marked with white bar.

![Figure 4](image_url)

Fig. 4. Example pattern of nine homozygous regions and interesting regions A1 and A2. The allele that is present in the reference sample r1 is shown in gray and the other allele is black. A1 consists of black alleles that are present in at least two overlapping regions, whereas the region A2 requires no allele specific overlapping as the reference set contains no homozygosity at that region. Alleles have been encoded so that A = AA, B = BB and X = AB. Recessive mutations may be found from the regions r3, r4 and r8 as they overlap A1. Similarly, regions r5, r7 and r9 may harbour recessive mutations at A2.

Score values belong to the range \([0,k]\) and are used as sorting criterion. In our case study, we observed that the score values are distributed according to the function \(a/\text{regionNumber}\). Parameter \(\text{regionNumber}\) is the index of the interesting region within a sorted list of scores, whereas \(a\) refers to a scaling factor that is calculated by minimizing the function \(\sum |\text{score} - (a/i)|\).

We calculate the saturation point of the scores utilizing the symmetry of the function \(1/x\). The score function \(a/\text{regionNumber}\) is scaled by an estimated factor of \(y_s = \max (\text{score})/\text{count (regions)}\) to make it symmetrical. The intersection of the symmetry axis \(\text{regionNumber} \times y_s\) and the score function represents the saturation point: \(y_s = \sqrt{a/y_s}\).

2.3 Annotations
Identification of regions whose score is above the saturation point is followed by the assessment of biological significance for the contents of these regions. The first step is to obtain identifications for the transcripts inside the regions. This step is followed by linking the transcripts to biomedical databases and highlighting genes that have been, for example, associated with other cancers or relevant biological processes.

The gene annotations for the interesting regions are done automatically. Currently, we have implemented gateways to Ensembl and SNPs3D databases. From the Ensembl database we query relevant annotations for the regions. First, the lengths of the genes and the exon counts can be used to estimate how much work is needed to sequence the possible mutations. Second, the microarray probe identifiers bring forward possibilities to integrate functional gene expression data. Third, we have developed a query tool (GO Search), which combines Ensembl gene annotations and the Gene Ontology (GO) database. GO Search selects interesting regions with genes having GO annotations that refer to a given GO term or any of its child terms. Fourth, the list of gene identifiers and descriptions can be compared against other databases. An example of this kind of interesting database is SNPs3D from which we can identify genes in the interesting regions that are associated with cancers or other diseases already. Disease Candidate Gene module of the SNPs3D database provides an interface to a text-mining tool that can be used to extract gene names based on keywords that are used in the abstracts of articles available in Medline. The keywords have been chosen based on their statistical support in abstracts containing the name of the disease in comparison to abstracts in general (Lin et al., 2004).

3 INHERITED COLORECTAL CANCER STUDY
We have applied CohortComparator to a proof-of-principle study where our objective is to detect recessively inherited colorectal cancer (CRC) predisposition loci. The risk of getting colorectal cancer is doubled if it has been diagnosed from a first degree relative and even four times the average if the relative was younger than 45 years old at the time of the diagnosis (Johns and Houlston, 2001). Many CRC predisposition genes are already known but based on epidemiological studies (Lichtenstein et al., 2000) many CRC families are likely to
segregate unknown dominant or recessive susceptibility alleles. The phenotypic effect of an ancient recessively inherited founder mutation can be seen in those who have inherited the mutation from their both parents that are descendants of the person who had the original mutation. This pattern of recessive inheritance is particularly likely in isolated populations such as Finns (Varilo et al., 2003).

3.1 Materials
Samples of the colorectal cancer patients (S\(D\)) originated from Eastern Finland. Total of 49 patients were selected out of 1044 prospectively collected colorectal cancer cases (the original collection of colorectal cancer patients is described in (Aaltonen et al., 1998; Salovaara et al., 2000)). Selection criterion was colorectal cancer in at least one sibling of a proband, and no evidence of dominant transmission. The cancers were microsatellite stable, and except for one sample, the common MYH gene mutations causing recessive CRC were excluded. The sample having MYH mutation was identified during the course of analysis. It was, however, kept in the patient cohort to highlight how CohortComparator and RegionAnnotator work.

Normal tissue DNA was extracted from frozen tissue or blood samples of the selected 50 patients and hybridized by Expression Analysis, Inc. (Durham, NC, US). The microarray experiments were carried over using Affymetrix GeneChip\textsuperscript{\textregistered}Human Mapping 100K Set and the standard protocol (Affymetrix, 2004). The 100K set actually consists of two 50K arrays (HindIII and XbaI). The 51 reference samples (S\(R\)) originated from different sources and were analysed using only XbaI array. The sample donors were not close relatives of each other.

Samples with a mean call rate \(<95\%\) were considered unsuccessful. Unsuccessful cases (8) were left out of the analysis as they would have introduced too many false positives in homozygosy analysis. Total of 41 patient samples were left for downstream analyses. The sample material was not collected from tumours but the surrounding healthy tissue. Therefore, our results are not biased due to large chromosomal aberrations that are common in colorectal cancers.

3.2 Haplotype estimation for the reference samples
The haplotype estimation of missing HindIII values for the reference samples was used to increase the minimum length of homozygous region in the reference set. The SNPs that were not measured from the reference samples were replaced with NoCall-values and all samples are analysed with the fastPHASE haplotyping method (Scheet and Stephens, 2006). Missing values in the SNP genotypes data were estimated based on the haplotype estimation they match the best. The haplotype estimation was conducted without pedigree information as the samples are unrelated. The haplotypes of the population were calculated from all samples. Before haplotyping, genomic positions matching the HindIII SNPs positions in the references samples were set to missing values. The mutation in interest is probably shared by only few individuals where the rest of the samples are ‘normal’. The alleles of the normal samples are more likely to bind the references as they share the common haplotype whereas the mutation is likely to differ in its XbaI constitution also. The computationally predicted HindIII values were converted back to SNP genotype data matrices with the full SNP resolution and the analysis was conducted using the original patient data parameters for the both data sets.

The results of the haplotype estimations were in better balance than the original results, when measured as the spatial densities of SNPs within homozygous regions along the chromosomes (results not shown). The reduction of the bias in their homozygosities can be explained as disappearance of many false positives that would have not appeared with a better SNP resolution. The chromosomal areas of interesting regions were quite well conserved between the original data and the haplotype estimates but the short fragments of the original data had fused into wider regions.

4 RESULTS
All 22 autosome were taken into account in CRC evaluation of CohortComparator. The top scored regions were inspected using the data visualization module to see how they are related to other regions and the samples. The genome-wide analysis resulted in 1874 interesting regions and after using the scoring method 181 regions were accepted covering total of \(\sim 137\) Mbp. No interesting regions were found from the chromosomes 19 and 22 as relatively few SNPs from these chromosomes were on the array. The low SNP density of telomere and centromere regions affected the results so that the number of homozygous regions found on these areas was less than expected.

4.1 Simulations
In this section, we provide two simulation cases to demonstrate the chances of detecting true homozygous regions and the effectiveness of our methodology to detect loci with recessive mutations. Before the actual simulations, we studied the effects of the length limits by comparing different combinations of \(l^D\) and \(l^R\) that represent 1 cases and 1 controls, respectively. Balance between the two exclusive parameters was found to match the surroundings of \(l^D = l^R = 55\). Further increase in length lead to lower sensitivity, while shorter lengths increased the amount of homozygous regions to such extend that the reference set filtered out most findings.

In order to examine the chance of getting homozygous regions without autozygosity, we conducted a simulation with randomly generated data. In this simulation, the frequencies of AA, AB and BB were first calculated for each SNP belonging to the chromosome 5. The frequencies were calculated from cancer patient samples and the distributions were then used to generate 51 random samples of chromosome 5. The SNP-specific genotype frequencies were used to weight the genotype probabilities and each SNP was assigned a value independently. The number of homozygous regions and their chromosome coverage were calculated with different values of \(l\) and the same analysis was completed with the original cancer patient data. The results are given in Figure 5. The number of homozygous regions found from the random samples decreased well below 0.05, when \(l = 54\) but an average of
approximately 4.67 homozygous regions were still found from the patient samples. The SNP genotypes of the real samples were clearly dependent on surrounding SNPs and distinctive from the independent genotypes of random data.

Analysis of the gap rate parameters revealed that the detection of homozygous regions is dominated by the parameters $l$ and $s$, if $g$ is small. The window length parameter $w$ was observed not to be critical for the results. Combinations of different $w:s$ for reference and patient data show only minor differences when $1/g \leq w \leq l$. The lower bound $w \geq 1/g$ allows at least one gap giving a motivation for the use of sliding window. Window sizes greater than $l$ are somewhat irrelevant since we are interested in shorter homozygous regions surrounded by heterozygous continuums. A too long window would not allow gap in such islands because the gap would be exhausted by the surroundings.

In the second simulation we studied how effective our methodology is detecting loci with recessive mutations, if such exist in the data. The simulation was based on the assumption that a genomic region can be either homozygous or not. This assumption allowed us to model the mutation detection probability with binomial distribution, which offered a framework to analytically study the effectiveness of our methodology. Using this framework we simulated different cohort sizes and different parameters for the probabilities of cases and controls (misclassified samples) to have a homozygous mutation. Our results demonstrate that CohortComparator is capable of detecting loci with recessive mutations, if such exist in the data, with a good probability. More detailed discussion on the simulation specifications and the results are given in the Supplementary Material.

4.2 Integration of the results to databases

After obtaining a gene list from computational analysis, the next step is to prioritize the genes for experimental validation. The nature of data and objectives more or less define which genes are considered as the most attractive gene candidates. Thus, the prioritization is usually done using various criteria, such as feasibility to genotype a gene, gene’s function, the assessment on gene’s impact on the disease progression and statistical significance. Sometimes we cannot expect that a list of genes emerging from the analysis would be statistically significant rendering statistical testing not an amenable option. Often, however, there is some intuition on biological processes that are likely to be associated with the disease progress, which can be used in conjunction with biological databases and literature to direct the search for the most attractive candidate genes for further experimentation. To facilitate such prioritization efforts, we have implemented tools to automatically query two biological databases.

In order to identify genes that are associated with cancers we use the SNPs3D database, which contains several methods to study gene-pathway-disease models (Lin et al., 2004). Briefly, annotations of the 181 interesting regions that exceeded the score limit revealed 1360 genes classified as Ensembl categories novel or known. The gene set was compared against the CRC genes listed in SNPs3D database. The comparison between the gene lists was completed by means of a simple text search capable of detecting gene names from the descriptions texts as well. The set of 27 genes were found to match the CRC gene list. An example of the results by RegionAnnotator is given in Table 1. Since we expect no over-representation of cancer genes in the samples, the SNPs3D list is not statistically significant (confirmed with random sampling test; results not shown).

Another way to prioritize genes is to find genes that are associated with a specific biological process. Such queries can be accomplished with the Gene Ontology database (Ashburner et al., 2000). An example of using the GO database is to compare the gene ontology annotations between the genes of all interesting regions and interesting regions exceeding the score
Identification of recessive mutations

Table 1. Colon cancer associated genes based on the SNPs3D annotations. SNPs3D reference column describes the name of the gene associated with colon cancer. The corresponding genes of interesting regions are listed after the SNPs3D entry.

<table>
<thead>
<tr>
<th>SNPs3D ref</th>
<th>Locus</th>
<th>Gene name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLH3</td>
<td>14: 74553239-74587886</td>
<td>MLH3</td>
<td>DNA mismatch repair protein Mlh3 (MutL protein homolog 3).</td>
</tr>
<tr>
<td>FHT</td>
<td>3: 59712992-60497735</td>
<td>FHT</td>
<td>Bis(5-adenosyl)-triphosphatase (EC 3.6.1.29) (Diadenosine 5',5'-P1,P3-triphosphate hydrolase) (Dinucleoside triphosphatase) (AP3A hydrolase) (AP3AASE) (Fragile histidine triad protein).</td>
</tr>
<tr>
<td>ERBB2</td>
<td>17: 35104766-35138441</td>
<td>ERBB2</td>
<td>Receptor tyrosine-protein kinase erbB-2 precursor (EC 2.7.10.1) (p185erbB2) (C-erbB-2) (NEU proto-oncogene) (Tyrosine kinase-type cell surface receptor HER2) (MLN 19).</td>
</tr>
<tr>
<td>MAPK4</td>
<td>17: 11868466-11987685</td>
<td>MAPK4</td>
<td>Dual specificity mitogen-activated protein kinase kinase 4 (EC 2.7.10.2) (MAP kinase kinase 4) (JNK-activating kinase 1) (c-Jun N-terminal kinase kinase 1) (JNKK) (SAPK/ERK kinase 1) (SEK1).</td>
</tr>
<tr>
<td>GCLC</td>
<td>6: 53470098-53517790</td>
<td>GCLC</td>
<td>Glutamate-cysteine ligase catalytic subunit (EC 6.3.2.2) (Gamma-glutamylcysteine synthetase) (Gamma-ECS) (GCS heavy chain).</td>
</tr>
<tr>
<td>PEX7</td>
<td>6: 137185410-137276752</td>
<td>PEX7</td>
<td>Peroxisomal targeting signal 2 receptor (PTS2 receptor) (Peroxin-7).</td>
</tr>
<tr>
<td>WNT5A</td>
<td>3: 55479912-55489996</td>
<td>WNT5A</td>
<td>Protein Wnt-5a precursor.</td>
</tr>
<tr>
<td>HMGBI</td>
<td>13: 29930884-30089729</td>
<td>HMGBI</td>
<td>High mobility group protein B1 (High mobility group protein 1) (HMG-1).</td>
</tr>
<tr>
<td>SRC</td>
<td>3: 36397101-36564500</td>
<td>SRC</td>
<td>SH3 and cysteine-rich domain-containing protein (SRC homology 3 and cysteine-rich domain protein).</td>
</tr>
<tr>
<td>SELE</td>
<td>1: 167958406-167968927</td>
<td>SELE</td>
<td>E-selectin precursor (Endothelial leukocyte adhesion molecule 1) (ELAM-1) (Leukocyte-endothelial cell adhesion molecule 2) (LECAM2) (CD62E antigen).</td>
</tr>
<tr>
<td>DVL3</td>
<td>14: 18535978-185374092</td>
<td>DVL3</td>
<td>Segment polarity protein dishevelled homolog DVL-3 (Dishevelled-3) (DSH homolog 3).</td>
</tr>
<tr>
<td>BDKRB1</td>
<td>14: 95799760-95800847</td>
<td>BDKRB1</td>
<td>B1 bradykinin receptor (BK-1 receptor) (B1R).</td>
</tr>
<tr>
<td>CCNG1</td>
<td>5: 162797155-162804598</td>
<td>CCNG1</td>
<td>Cyclin-G1 (Cyclin-G).</td>
</tr>
<tr>
<td>PLXDC1</td>
<td>17: 34473083-34561298</td>
<td>PLXDC1</td>
<td>Plexin domain-containing protein 1 precursor (Tumor endothelial marker 7) (Tumor endothelial marker 3).</td>
</tr>
<tr>
<td>ANTXR1</td>
<td>2: 69093993-69329804</td>
<td>ANTXR1</td>
<td>Anthrax toxin receptor 1 precursor (Tumor endothelial marker 8).</td>
</tr>
<tr>
<td>VCP</td>
<td>9: 35046061-35063246</td>
<td>VCP</td>
<td>Transitional endoplasmic reticulum ATPase (TER ATPase) (15S Mg(2+)- ATPase p97 subunit) (Valosin-containing protein) (VCP).</td>
</tr>
<tr>
<td>RPL31</td>
<td>2: 100985183-100989312</td>
<td>RPL31</td>
<td>60S ribosomal protein L31.</td>
</tr>
<tr>
<td>TGFA</td>
<td>2: 70527927-70634438</td>
<td>TGFA</td>
<td>Transforming growth factor alpha precursor (TGFA) (EGF-like TGF) (ETGF) (TGF type 1).</td>
</tr>
<tr>
<td>PPP1R1B</td>
<td>17: 35036705-35046403</td>
<td>PPP1R1B</td>
<td>Dopaamine- and cAMP-regulated neuronal phosphoprotein (DARPP-32).</td>
</tr>
<tr>
<td>ALOX5</td>
<td>10: 45189635-45216567</td>
<td>ALOX5</td>
<td>Arachidonate 5-lypoxgenase (EC 1.13.11.34) (5-lypoxgenase) (5-LO).</td>
</tr>
<tr>
<td>PLXDC2</td>
<td>10: 20145174-20609292</td>
<td>PLXDC2</td>
<td>Plexin domain-containing protein 2 precursor (Tumor endothelial marker 7-related protein).</td>
</tr>
<tr>
<td>TPBG</td>
<td>8: 83130067-83137264</td>
<td>TPBG</td>
<td>Trophoblast glycoprotein precursor (ST4 oncotrophoblast glycoprotein) (ST4 oncotrophoblast glycoprotein) (ST4 oncotrophoblast antigen) (M6P1).</td>
</tr>
<tr>
<td>UBA1P</td>
<td>9: 34169003-34242521</td>
<td>UBA1P</td>
<td>Ubiquitin-associated protein 1 (UBAP).</td>
</tr>
<tr>
<td>FOS</td>
<td>14: 74815284-74816865</td>
<td>FOS</td>
<td>Proto-oncogene protein c-fos (Cellular oncogene fos) (G0/G1 switch regulatory protein 7).</td>
</tr>
<tr>
<td>COP5</td>
<td>8: 68117871-68136785</td>
<td>COP5</td>
<td>COP9 signalosome complex subunit 5 (EC 3.4.-.-) (Signalosome sub- unit 5) (SGN5) (Jun activation domain-binding protein 1).</td>
</tr>
</tbody>
</table>

limit. The comparison between these genes was performed with High-Throughput GoMiner (Zeeberg et al., 2005) and the results suggested high enrichment of genes related to apoptosis and programmed cell death. Another example is to identify genes that were associated with apoptosis, and 30 interesting regions exceeding the score limit were found. The regions contained total of 44 apoptosis-related genes. Interestingly, a single region in chromosome 14 contained four genes: CIDEB, GZMB, GZMH and RIPK3.

Single sample having MYH mutation in the set of 41 samples was not identified, most likely because the sample size of 41 is too small to detect proportion of 2.4%. Thus, we used XbaI array to measure 50 K SNPs from an additional sample with a known homozygous MYH mutation different from the one originally included in the study (Enholm et al., 2003). The additional sample was analyzed together with the 41 CRC and 50 reference samples (details in Supplementary Material). After adding one sample with the MYH mutation, CohortComparator identified corresponding region accurately and further annotation with RegionAnnotator highlighted MYH as a top candidate gene.

The region annotations were also compared to the CRC candidate cancer genes listed in (Sjöblom et al., 2006). The article describes genes with observed mutations in colorectal and breast cancers and the candidate cancer genes are the ones with a mutation rate higher than the background rate. The mutations
in these genes may have contributed to the development of the tumour. Two genes, GNAS and GUCY1A2, were included in the top 181 scored homozygous regions.

5 DISCUSSION

We have developed a framework for genome-wide analysis and visualization of two cohorts that are subjected to SNP-microarray experiments. The main objective of the framework is to generate hypotheses that can be tested in future studies. For example, with increased genotyping throughput, studies focusing on identifying association between gene variations and a phenotype now face the problem of what SNPs to include in the analysis. Further, often a laboratory conducting such research already has SNP-array data sets that are hoped to direct the choice of genes to be genotyped in thousands of samples. The framework presented here provides means to integrate SNP-array data into chromosome-specific graphs and highlighting the data. CohortComparator takes an approach of splitting the human genome may hinder effective visualization of the samples. High rates of homozygosity may indicate familial relatedness of individuals’ parents. The relative amounts of homozygosity are readily shown in CohortComparator graphs, which help in estimating the distribution. The extensive size of the human genome may hinder effective visualization of the data. CohortComparator takes an approach of splitting the data into chromosome-specific graphs and highlighting the possibly interesting sample features based on their rarity in references. The clear separation between long homozygous regions and homozygous SNPs helps to focus on significant features of the data. The current version of CohortComparator is capable of detecting regions that could harbour compound heterozygosity. That is, samples that do not have SNPs of complement homozygosity to the interesting homozygous region. More sophisticated mechanisms can be achieved with the aid of haplotype data and this is one of our future directions to improve CohortComparator.

The selection of proper reference samples is a crucial step as the comparison between the data sets is extremely sensitive to references sharing the haplotype in interest. The colorectal study was completed with zero tolerance for homozygous haplotypes, which means that a single reference sample with a latent susceptibility would cover the site of mutation. The tolerance limit can be loosened as more reference samples become available, which should facilitate revealing the recessive mutations.

The choice of the parameters is an important step in explorative analysis. Here, we have given guidelines to tune the parameters in CohortComparator in order to obtain biologically feasible results. Consequently, in our case study the homozygous regions were long enough to be biologically feasible. Still, we also detected short regions consisting of only single genes. In future, the issue of the uneven distribution of measured SNPs along the chromatin could be compensated using an optional setup, where the parameters l and w are given in base pairs instead of SNPs.

In addition to homozygous regions, we identified 13 sites of possible deletions. All sites were found to be false positives when examined with the polymerase chain reaction essay. This result suggests that Affymetrix SNP-arrays may not be the best alternative for search for homozygous deletions as missing signals, at least in our data, seem to be due to technical artefacts rather than real overlapping deletions in both chromosomes. A way to overcome this technical problem could be to integrate copy number analysis that is based on the original probe intensities to haplotype data as described in Zhao et al. (2004).

CohortComparator provides means to link genome-wide gene variation results to functional gene expression data. For each gene in the interesting regions we automatically get gene ID and Affymetrix probe ID, which can be used to testing how these genes are expressed in a gene microarray study, or a microarray expression databases. A gene that is identified by CohortComparator and then found to be underexpressed is a very strong tumour suppressor gene candidate. In summary, CohortComparator and RegionAnnotator provide means to speed up the process to transform high-throughput SNP data set to biomedical knowledge.

ACKNOWLEDGEMENTS

Reference samples were kindly donated by Finnish Red Cross. We thank Finnish Genome Center and Dr Janna Saarela for technical help to get the data, and Scientific Computing Ltd. (CSC) for computing resources. Financial support from Academy of Finland, Biocentrum Helsinki, Finnish Cancer Organisations, the Sigrid Jusélius Foundation and University’s Research Funds is gratefully acknowledged.
Conflict of Interest: none declared.

REFERENCES


Yue, P. et al. (2006) SNPs3D: Candidate gene and SNP selection for association studies. BMC Bioinformatics, 7, 166.
