SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels
Mathijs A Sanders¹, Roel GW Verhaak¹,², Wendy MC Geertsma-Kleinekoort¹, Saman Abbas¹, Sebastiaan Horsman³, Peter J van der Spek³, Bob Löwenberg¹ and Peter JM Valk*¹

Address: ¹Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands, ²Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA; The Broad Institute of M.I.T. and Harvard, Cambridge, USA and ³Department of Bioinformatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Email: Mathijs A Sanders - m.sanders@erasmusmc.nl; Roel GW Verhaak - verhaak@broad.mit.edu; Wendy MC Geertsma-Kleinekoort - w.geertsma-kleinekoort@erasmusmc.nl; Saman Abbas - s.abbas@erasmusmc.nl; Sebastiaan Horsman - s.horsman@erasmusmc.nl; Peter J van der Spek - p.vanderspek@erasmusmc.nl; Bob Löwenberg - b.lowenberg@erasmusmc.nl; Peter JM Valk* - p.valk@erasmusmc.nl

* Corresponding author

Abstract

Background: Accurate analyses of comprehensive genome-wide SNP genotyping and gene expression data sets is challenging for many researchers. In fact, obtaining an integrated view of both large scale SNP genotyping and gene expression is currently complicated since only a limited number of appropriate software tools are available.

Results: We present SNPExpress, a software tool to accurately analyze Affymetrix and Illumina SNP genotype calls, copy numbers, polymorphic copy number variations (CNVs) and Affymetrix gene expression in a combinatorial and efficient way. In addition, SNPExpress allows concurrent interpretation of these items with Hidden-Markov Model (HMM) inferred Loss-of-Heterozygosity (LOH)- and copy number regions.

Conclusion: The combined analyses with the easily accessible software tool SNPExpress will not only facilitate the recognition of recurrent genetic lesions, but also the identification of critical pathogenic genes.

Background

High-density genome-wide views of biological samples, using high-throughput DNA mapping and mRNA gene expression microarrays facilitate the identification of recurrent molecular lesions. Both types of microarrays, which are being produced by different manufacturers, e.g., Nimblegen, Agilent, Sequenom, Applied Biosystems, Illumina and Affymetrix, typically contain large numbers of small oligonucleotides that interrogate the genome. Currently available DNA arrays contain over 500.000 probe sets, while the gene expression arrays target over 20.000 genes. Efficient analysis of these large datasets remains a challenge for many researchers.

The Affymetrix and Illumina DNA mapping platforms have been designed to specifically target sequences containing single nucleotide polymorphisms (SNPs). SNPs are currently estimated to be present at a frequency of 1
out of 300 nucleotides [1]. By including different probe sets to detect the possible SNP variants, genome-wide genotyping is feasible. In fact, these types of arrays have been developed for genome-wide association studies; however, these platforms can easily be applied to determine copy numbers of these chromosomal markers, similar to array comparative genomic hybridization (CGH). Because of the high number of SNPs, sample DNA can be examined with an inter-marker distance of 6 to 12 kb, and (micro) deletions and/or amplifications are detectable. By comparing disease samples to normal germ line DNA, a detailed overview of acquired gains and losses of the genome is obtained. In fact, although our knowledge is still developing, it has recently become apparent that that copy number variation (CNV) accounts for a substantial amount of genetic variation in the human genome [2]. The high-resolution scanning technologies enable the analyses of CNV and associated phenotypes [2].

The power of DNA mapping has been shown extensively in cancer research. Chromosomal gains and losses as well as regions of loss-of-heterozygosity (LOH) have been shown in, for instance, leukemia [3,4], lung cancer [5-7] and colon cancer [8]. Recognition of recurrent lesions will ultimately result in the identification of pathogenic genes. For instance, SNP array analysis of a set of cancer cell lines has lead to the identification of the microphthalmia-associated transcription factor MITF as a melanoma oncogene [9].

On the Illumina platform genotypes are determined using hybridization of genomic DNA to BeadChips followed by an enzymatic discrimination step. On the Affymetrix platform, genotype calls and copy numbers are determined by a probe set consisting of mismatch and perfect match probes. In analogy with the expression probe set, the genotype- and gene expression files required for SNPExpress can be generated through tools such as Affymetrix BRLMM [13], GCOS/CNAT 4.0 [24], or dChipSNP [17] with additional formatting in Microsoft Excel. In case of Illumina data, SNP Express includes the non-synonymous SNPs and the MHC region, however, mitochondrial SNPs and Y-chromosome SNPs are not visualized. All files can be optionally uploaded as tab- or comma-delimited .txt files or binary files. These binary files can be created from .txt files by the menu item ‘convert data source’.

SNPExpress maps both the SNP IDs (Illumina and Affymetrix) and the expression probe set IDs (Affymetrix) to the genome through internal alignment tables, using annotation provided by the manufacturer [25,26] and [27]. Annotation was generated using NCBI build 36.1.

Regions showing LOH are calculated through a hidden Markov Model, which has been described previously [18]. The probability values for heterogeneous calls required for the HMM have been generated through sets of genotypes of normal samples. For the 100 K and 500 K array, 90 samples and 270 samples, respectively, of different ethnic background from the HapMap project are available through the NCBI GEO website (and provided by the
and its associated SNPs are indicated with an orange background. The options to display known CNVs (purple background) or the HMM copy number results (thin magenta line) are included (Figure 1C). Finally, relevant data of a particular minimal deleted of amplified region can be exported (i.e. Sample, Probe_set_id, Chromosome, Location (bp), Cytoband, Associated gene, Genotype, Copy number and Inferred LOH of the selected region) and high-resolution images of the visualization can be saved in the Portable Network Graphic (PNG) format.

To illustrate the power of SNPExpress, DNA mapping array profiles of tumor samples of a series of 48 patients with AML were generated using Affymetrix 250 K NspI DNA mapping arrays. Ficoll separation of the mononuclear cells from AML typically yields >80% pure population of leukemic blast cells. High molecular weight DNA was isolated from these malignant cells and the Affymetrix mapping arrays were used according to the protocol of the manufacturer. Genotypes were calculated using BRLMM and copy numbers were assessed using dChipSNP. Biotin-labeled cRNA of the same AML samples was hybridized on Affymetrix HG-U133 plus 2.0 GeneChips, as described elsewhere [34]. The resulting dataset was imported in SNPExpress for analyses. Large chromosomal regions showing loss or gains of genetic material are known to be apparent in leukemic blasts of AML patients. Well-known examples of chromosomal lesions in AML are monosomies of chromosome 5 and 7, which have been associated with a poor prognosis [35]. Using SNPExpress, monosomies of chromosome 7 were evidently demonstrated in AML samples, previously shown by cytogenetics to have lesions involving chromosome 7 (Figure 1). SNPExpress also correctly predicted the presence of LOH as a result of the absence of one chromosome 7. In fact, 17 out of 21 numerical cytogenetic aberrations, i.e., whole chromosome losses and interstitial deletions, in 48 AML samples analyzed, were recognized by using SNPExpress. Four numerical abnormalities abnormalities, present in less than 30% of the AML cells, were missed. Chromosomal gains, losses as well as uniparental disomy (UPD) may also have other important consequences, such as affecting expression of (imprinted) genes. Combinatorial visualization of genotype, copy number and gene expression is a prerequisite to recognize these aberrations. For example, the majority of genes show located on chromosome 7 show an overall decrease in expression in AML samples with a monosomy 7 (Figure 1).

Large regions of homozygosity are present in approximately 20% of primary AML cases as a result of segmental UPD [3,36]. These regions of UPD seemed to be non-random and may be used to unmask pre-existing recessive mutations in leukemia genes, such as CEBPA, WT1, FLT3 and RUNX1 [3,37]. SNPExpress adequately identified regions of UPD involving e.g. chromosome 11p (Figure 1). Example expression, copy number, genotype and HMM copy number example files of two AML patients can be downloaded from [33].
1D), in two patients with a normal karyotype. UPD involving chromosome 11 is associated with homozygous mutations in WT1 [37]. Interestingly, in 13 out of 48 AML patients (27%) large regions of segmental UPD continuing to the telomere were recognized using SNPExpress. These examples demonstrate the power of SNPExpress. To our knowledge, no tool is currently available that allows concurrent interpretation of genotype, HMM inferred LOH regions, copy number, CNVs, HMM inferred copy

Figure 1
SNPExpress Screenshot. A. DNA mapping array data from the Affymetrix 250 K NspI DNA mapping array was used to sequentially align the genotypes and copy numbers of chromosome 7 of four AML samples. The copy numbers (n = 0, 1, 2, 3, 4) are shown for each individual patient by horizontal lines. Copy number n = 2 is depicted by a green line (A). The SNP genotypes are sequentially aligned along the chromosome (AA: red; BB: yellow; AB: blue, noCall: white). LOH is indicated by a thick magenta horizontal bar (B), gains (default n > 2.5) by a pink (Figure 1C) and losses (default n < 1.5) by a turquoise background (C). Gene expression levels are visualized as vertical white bar at the chromosomal position of the gene-specific probe set. In the event that multiple probe sets span the same region in the chromosome-wide view the vertical gene expression bars are red and proportional to the highest expression value. The two upper samples clearly display a decreased copy number as was previously shown by cytogenetics, i.e., a complete monosomy (sample 1) or a deletion of the q-arm of chromosome 7 (sample 2). The overall expression of the majority of genes in the displayed region is decreased in the samples with chromosome 7 abnormalities. The chromosome selector (D; where 23 is the X chromosome), the mouse-over function showing info of each SNP or probe set (E), full chromosome view (F), zoom function (G) gene search function (H), the links to external databases (I), display CNVs (J) and export selected data (K) options are indicated. B. Full chromosome view of samples from 1A. C. CNV (purple background) and copy number of each SNP based on a HMM model (HMM copy number, magenta line) of the two AML patients from examples [33]. In the event that multiple CNVs span the same region in the chromosome-wide view the background is violet, whereas single CNV are indicated with a rosy brown background. D. UPD of chromosome 11 demonstrated using SNPExpress. Example of large scale UPD on chromosome 11 in the upper two AML patients with a normal karyotype in comparison to two other AML samples. The overall copy number is two and large regions of LOH are indicated by the thick magenta line across the chromosome. After using the search function, SNPs associated with WT1 are depicted with an orange background.
number and gene expression data. Moreover, no specialized knowledge is necessary to work with SNPExpress.

Discussion
Since genome-wide DNA mapping array and mRNA expression studies become more cost effective, the number of samples profiled on these platforms will increase. Specialized user-friendly tools for efficient visualization, such as SNPExpress, will therefore be indispensable. In fact, the initial version of SNPExpress has already been successfully applied in showing segmental uniparental disomy as a recurrent mechanism for homozygous CEBPA mutations in acute myeloid leukemia [38].

Other tools for visualizing and processing SNP array data, such as SNPScan [39], SIGMA [40], Array Fusion [41], Partek Genomics Suite [42] and GenePattern [43] have been developed. Most of these tools incorporate visualization options for displaying LOH (GenePattern, Partek Genomics Suite, SNPScan) and copy number (all but ArrayFusion), whereas SNPScan and ArrayFusion have output functionality that facilitates linking SNP data to the UCSC genome browser [39,41]. Some are linked to a private database, which restricts pre-processing of the array data, but gives the advantage of data storage [40]. GenePattern and the Partek Genomics Suite provide normalization and data smoothing functionality. These two packages and SNPScan have also incorporated options for combined analysis of paired samples, i.e., tumor and normal. Like SNPExpress, SNPscan, GenePattern, and the Partek Genomics Suite can detect regions of LOH, amplification and deletion. None of these tools describe the ability to process Illumina BeadArray files. Where SNPExpress may lack the opportunity to directly process raw data files (such as Affymetrix CEL-files), it adds integrated visualization of expression (Affymetrix) and DNA copy number and genotype (Affymetrix and Illumina) data. Moreover, we believe that this is provided in a user-friendly way that does not require specialist computer knowledge.

SNPExpress has some limitations. A full-length chromosome view depicting gains, losses and the regions showing LOH is feasible using SNPExpress. However, the large datasets generated by the 500 K mapping array platform makes it impossible to visualize the sequentially aligned SNPs of the full-length chromosomes on one screen. Selecting the most informative SNPs, i.e., representative for particular haplotypes, may solve this issue. Such algorithms are currently in development. Furthermore, the current implementation of the HMM could also be improved by implementing a HMM that takes into account the effects of linkage disequilibrium, i.e., LD-HMM [18]. The number of samples to be visualized currently is limited by the memory available to the application.

Conclusion
The power of SNPExpress, as with previously developed tools [44], is its high accessibility and powerful visualization, which facilitates the identification of biologically and clinically relevant entities. We have shown that recurrent biologically relevant entities, such as chromosomal gains or losses and LOH in AML, are accurately identified with SNPExpress. Hence, SNPExpress will be beneficial to genome-wide studies by providing an integrated view of data from DNA mapping and mRNA expression arrays in an easily accessible and accurate way.

Availability and Requirements
Project name: SNPExpress

Project homepage: http://www.erasasmusmc.nl/hematologie/SNPExpress

(Including downloadable genotype-, copy number-, expression- and HMM copy number example files of two AML patients genotyped with Affymetrix 250 K NspI DNA mapping array and gene expression profiled with Affymetrix U133Plus2.0 GeneChips)

Operating system: Platform independent

Programming language: JAVA

Other requirements: JAVA 1.5 or higher.

License: The tool is available free of charge. Source code is available upon request.

Any restrictions to use by non-academics: None

List of Abbreviations
AML: Acute Myeloid Leukemia; PNG: Portable Network Graphics; BRLMM: Bayesian robust linear model with Mahalanobis distance classifier; SNP: Single nucleotide polymorphism; HMM: Hidden Markov Model; CNV: Copy Number Variation; LOH: Loss-of-heterozygosity.

Authors’ contributions
MAS wrote and designed the software; RGWV designed the software, performed the analysis and wrote the manuscript; WGK performed experiments; SA gave intellectual contributions; SH contributed code; PJS gave intellectual contributions; BL gave intellectual contributions; PJMV designed the study, gave intellectual contributions and wrote the manuscript. All authors read and approved the final version of the manuscript.
Acknowledgements
The research described was supported by grants from the Erasmus University Medical Center (Revolving Fund) and the Dutch Cancer Society "Konin gin Wilhelmina Fonds". We are indebted to Andy Hall for providing Affymetrix 10 K DNA mapping array data at the initial set up of SNPExpress.

References
1. International HapMap Consortium: A haplotype map of the human genome. Nature 2005, 437(7063):1299-1320.
2. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Abtahian SH, Abecasis GR, Agarwal SK, An H, Antonacci F, Musunuru K, Babu MS, Chang JF, Colegrave RN, Corbeil J, Cottrell GC, DeBol t DC, Dunning M, Eichler EE, E早晨mon JH, Erskine A, Fraga MF, Gilad S,虢 S, Green RM, Gyongyosi B, Heid IM, House MJ, van der Spek PJ, Löwenberg B, Delwel R: Copy number detection using high-density oligonucleotide microarrays. Genome Res 2005, 15(11):1717-1724.
3. Raghavan M, Lillington DM, Skoulakis S, Debernardi S, Chaplin T, Foot PP, House MJ, van der Spek PJ, Löwenberg B, Delwel R: Frequent loss of heterozygosity without loss of genetic copy number in colorectal cancer. Cancer Res 2005, 65(13):5561-5570.
4. Irving JA, Bloodworth L, Bown NP, Case MC, Hogarth LA, Hall AG: Identification of LOH in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. Cancer Res 2005, 65(8):3033-3058.
5. Zhao X, Wei BA, Laframboise T, Lin M, Beroukhim R, Garraway LA, Beheshti J, Lee JC, Naoki K, Richards WG, Sugarbaker D, Chen F, Rubin MA, Janne PA, Girard L, Minna J, Christiani DC, Johnson BE, Lander ES, Meyerson M: Loss-of-heterozygosity and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis. Cancer Res 2005, 65(13):5561-5570.
6. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Liu WQ, Villapakka M, Stanton A, Larson C, Hudson TJ, Johnson BE, Lander ES, Meyerson M: Copy number variations in human cancer genomes. Nature Genet 2005, 37(5):248-254.
7. Raghavan M, Lillington DM, Skoulakis S, Debernardi S, Chaplin T, Foot PP, House MJ, van der Spek PJ, Löwenberg B, Delwel R: Frequent loss of heterozygosity without loss of genetic copy number in colorectal cancer. Cancer Res 2005, 65(13):5561-5570.
tion in acute myeloid leukemias. Cancer Res 2005, 65(20):9152-9154.

38. Wouters BJ, Sanders MA, Lugthart S, Geertsma-Kleinekoort WMC, van Drunen E, Beverloo HB, Lowenberg B, Valk PJM, Delwel R: Segmental uniparental disomy as a recurrent mechanism for homozygous CEBPA mutations in acute myeloid leukemia. Leukemia 2007, 21(11):2382-4.

39. Ting JC, Ye Y, Thomas GH, Ruczinski I, Pevsner J: Analysis and visualization of chromosomal abnormalities in SNP data with SNPscan. BMC Bioinformatics 2006, 7(1):25.

40. Chari R, Lockwood WW, Coe BP, Chu A, Macey D, Thomson A, Davies JJ, MacAulay C, Lam WL: SIGMA: a system for integrative genomic microarray analysis of cancer genomes. BMC Genomics 2006, 7:324.

41. Yang TP, Chang TY, Lin CH, Hsu MT, Wang HW: ArrayFusion: a web application for multi-dimensional analysis of CGH, SNP and microarray data. Bioinformatics 2006, 22(21):2697-8.

42. Partek Discovery Suite [http://www.partek.com]

43. Reich M, Lifeld T, Gould J, Lerner J, Tamayo P, Mesirov JP: GenePattern 2.0. Nat Genet 2006, 38(5):500-1.

44. Verhaak RG, Sanders MA, Bijl MA, Delwel R, Horsman S, Moorhouse MJ, van der Spek PJ, Lowenberg B, Valk PJ: HeatMapper: powerful combined visualization of gene expression profile correlations, genotypes, phenotypes and sample characteristics. BMC Bioinformatics 2006, 7:337.