# Mutational Genomics for Cancer Pathway Discovery

Jeroen de Ridder[1,2,3], Jaap Kool[4,6], Anthony G. Uren[5,6], Jan Bot[1,3], Johann de Jong[2], Alistair G. Rust[7], Anton Berns[6], Maarten van Lohuizen[6], David J. Adams[7], Lodewyk Wessels[1,2,3], and Marcel J.T. Reinders[1,3]

[1] Delft Bioinformatics Lab, Delft University of Technology
[2] Bioinformatics and Statistics, Dept. Molecular Biology, Netherlands Cancer Institute
[3] Netherlands Bioinformatics Centre
[4] MSD Animal Health, Merck/Intervet B.V.
[5] MRC Clinical Sciences Centre, Imperial College Faculty of Medicine
[6] Dept. Molecular Genetics, Netherlands Cancer Institute
[7] Experimental Cancer Genetics, Wellcome Trust Sanger Institute
l.wessels@nki.nl, m.j.t.reinders@tudelft.nl

**Abstract.** We propose *mutational genomics* as an approach for identifying putative cancer pathways. This approach relies on expression profiling tumors that are induced by retroviral insertional mutagenesis. Akin to genetical genomics, this provides the opportunity to search for associations between tumor-initiating events (the viral insertion sites) and the consequent transcription changes, thus revealing putative regulatory interactions. An important advantage is that in mutational genomics the selective pressure exerted by the tumor growth is exploited to yield a relatively small number of loci that are likely to be causal for tumor formation. This is unlike genetical genomics which relies on the natural occurring genetic variation between samples to reveal the effects of a locus on gene expression.

We performed mutational genomics using a set of 97 lymphoma from mice presenting with splenomegaly. This identified several known as well as novel interactions, including many known targets of *Notch1* and *Gfi1*. In addition to direct one-to-one associations, many multilocus networks of association were found. This is indicative of the fact that a cell has many parallel possibilities in which it can reach a state of uncontrolled proliferation. One of the identified networks suggests that *Zmiz1* functions upstream of *Notch1*. Taken together, our results illustrate the potential of mutational genomics as a powerful approach to dissect the regulatory pathways of cancer.

## 1 Introduction

Cancers arise as a result of a multistep process in which genetic alterations deregulate the regulatory pathways that govern healthy cell proliferation [1]. To study this process, the use of DNA microarrays for transcriptome profiling of tumor tissue has proven useful. Success stories include, among others, finding good diagnostic and prognostic markers [2, 3], and providing insight in different tumor subtypes [4]. However, to identify the *causal* genetic alterations, transcriptome profiling is less suitable. This is because, in many cases, aberrant gene expression is a downstream effect of one or more genetic alterations elsewhere, rather than the causal event in tumor development.

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 35–46, 2013.

To identify genes that are likely to have a driving role in cancer, high-throughput retroviral insertional mutagenesis (RIM) screens can be performed [5–8]. In these screens, retroviruses are used to induce insertion mutations in the genome of infected somatic cells in mice. These mutations may cause alteration in expression of genes in the vicinity of the insertion or, when inserted within a gene, alteration of the gene product. A certain proportion of these mutations are oncogenic and will result in tumor development. Consequently, the genomic location of the inserted viruses in the resulting tumors provide 'tags' for cancer genes, since regions in the genome that harbor insertions in multiple independent tumors are likely to be in the vicinity of genes that play a causal role in tumor development.

## 1.1 Mutational Genomics

Here, we perform genome-wide expression profiling in tumors induced by RIM. Combining expression with insertion site data provides the unique opportunity to study the relationship between the initiating events and their downstream transcriptional effect. We call this approach *mutational genomics*.

Mutational genomics bears similarity to genetical genomics, linking genotype to transcriptional state [9–11]. In the latter approach, often performed in fully genotyped recombinant inbred (RI) mouse strains, expression quantitative trait loci (eQTLs) are determined. These are defined as chromosomal regions for which the local genotype segregates the gene expression of one or more genes, and may point to putative regulators of these genes [12–14]. Similarly, mutational genomics allows the definition of, what we coin, expression quantitative mutation loci (eQMLs), i.e. chromosomal regions that are mutated in multiple independent tumors and are associated with a segregation of the expression of one or more genes. This concept is schematically illustrated in Figure 1.

A major advantage of mutational genomics is that the list of candidate target genes of the identified eQMLs is usually limited to only a few. This is because insertions act primarily on proximal genes [15] using one of a specific set of fairly well defined mechanisms [5, 7, 16]. Typical eQTLs, on the other hand, usually span large regions in the genome containing many genes as a result of linkage disequilibrium. Consequently, in mutational genomics the difficult task of finding the genes underlying the transcriptional changes is circumvented.

A second important advantage stems from the fact that mutational genomics exploits the selective pressure exerted during tumor development to yield a relatively small number of loci that are likely to be causal for tumor formation. This is unlike genetical genomics in which one has to rely on the natural occurring genetic variation between samples to reveal the effects of a locus on gene expression. As a result, eQMLs are specific for the type of tumor under study, and therefore represent important building blocks that help delineating the regulatory pathways that play a role in these tumors.

## 1.2 Multilocus Interactions

Cancer is a complex disease, involving the mutation and/or deregulation of multiple genes. Many of the changes that are required for tumorigenesis are a result of the collaboration between mutations of cancer genes. Moreover, for many of the mutational steps

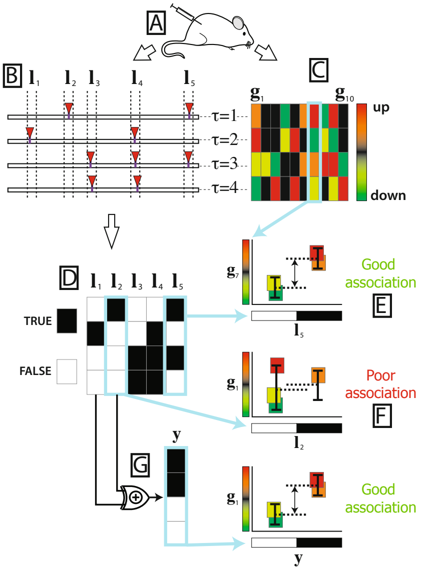


**Fig. 1. Schematic overview of the data for four tumors. A)** After infection with a slow transforming retrovirus, tumors are harvested. **B**) The insertion loci are retrieved by sequencing the flanking regions. The figure shows five unique insertion loci ($\mathbf{l}_1 - \mathbf{l}_5$), for four tumors ($\tau = 1, \ldots, 4$). **C**) For each tumor, gene expression profiles are determined by microarrays. The figure shows 10 genes ($\mathbf{g}_1 - \mathbf{g}_{10}$). **D**) The insertion data can be considered as a Boolean matrix. **E**) An insertion locus is said to be associated with the expression of a gene when the presence or absence of an insert segregates the gene expression in a highly expressed and lowly expressed group, as is the case for inserts in $\mathbf{l}_5$ and expression of $\mathbf{g}_7$. **F**) In some cases a single insertion locus does not suffice to explain the expression values, exemplified by the poor association between $\mathbf{l}_2$ and $\mathbf{g}_1$. **G**) Multilocus models, combining multiple loci using Boolean logic ($\mathbf{l}_1$ XOR $\mathbf{l}_2$), may be employed to explain more of the transcriptional variance.

required to transform healthy cells to cancer cells numerous alternatives exist. This is especially pertinent while analyzing mutational genomics data, since this means that many of the regulatory interactions may not be detectable as direct (marginal) associations, but rather require multivariate analysis of the data (see Figure 1G for a schematic example).

Therefore we propose to explore multilocus mapping by explicitly incorporating the possibility of alternative and collaborative pathways in the search for eQMLs. Because the presence or absence of an insertion is naturally captured by a Boolean variable, a Boolean model is used to combine insertion loci. To this end, we employ the combinatorial association logic (CAL) network inference procedure, that we recently proposed for finding multilocus interaction in a genetical genomics dataset [17]. Using CAL network inference we are able to efficiently determine the set of insertion loci that, when combined using a Boolean logic function, shows strong association with the gene expression levels.

## 2 Results

We have performed Mutational Genomics of a set of 97 retrovirally induced splenic lymphomas in p19$^{\text{ARF}-/-}$ ($n$=31), p53$^{-/-}$ ($n$=19) and wt ($n$=53) mice. The retroviral insertion sites found in these tumors have been published previously[1] [18]. Gene expression data were obtained using the Illumina MouseWG6-V2 beadchips. A detailed

[1] Available at `http://mutapedia.nki.nl`

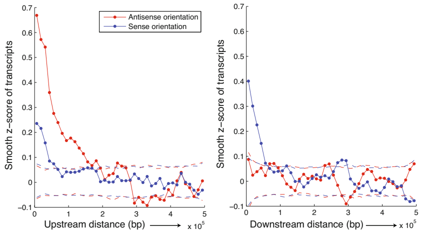


**Fig. 2. Insertion alignment plots showing effect of insertions on transcription.** The solid lines represent the smoothed z-scores of transcripts with insertions upstream (left) or downstream (right). Distance is relative to transcription start sites. Insertions were also split according to their orientation relative to the transcripts with red lines indicating 'anti-sense' insertion effects (insertion orientation opposite to transcript orientation). The inverse holds for the lines. The dashed lines reflect the 5% significance threshold, obtained by permutation.

description of the preprocessing of the data can be found in the Methods section and the Supplementary material.

## 2.1 Insertions Affect Local Transcription

We first investigated the local effect of the insertional mutations on transcription. Figure 2 shows a genome-wide alignment of all insertions in the dataset. A point in this figure at $(d, z)$ represents the average $z$-score $(z)$ of the expression of all genes in a bin $d$ basepairs removed from the insertion. Panel A and B show the result for genes with upstream and downstream insertions, respectively, with different colors indicating insertion orientation relative to the transcript.

Figure 2 reveals that, on a global level, a clear effect of the insertions on the local transcription is present but that this effect is dependent on distance. Furthermore, it can be seen that antisense insertions result in a higher average expression, indicating a strong effect on local transcription, when their relative position to the transcript is upstream. Conversely, sense insertions seem to have a stronger effect in case they are positioned downstream of the transcript. These observations are consistent with previously described mechanisms through which retroviruses act on their targets [5, 7, 16]. For this reason we decided to implement a set of literature derived rules that map insertions to their putative target transcripts based on their relative position and orientation (see Supplements for details). This provides a mapping of all insertions in a given genomic locus to a unique identifier.

## 2.2 Mutational Genomics Reveals eQMLs

**Association Inference.** After normalization and selection of the most highly variable probes, probes were hierarchically clustered using a stringent correlation distance cutoff. This yielded 6228 clusters, henceforth referred to as gene clusters. For gene clusters containing multiple genes (1177 cases) cluster centroids were determined by taking the mean across the expression profiles.

To determine the Boolean insertion matrix (representing the insertion loci, see Figure 1D), all insertions were mapped to their target transcripts according to the literature derived rules. Each transcript represents one column of the Boolean insertion matrix and is determined by recording TRUE in case a tumor contains a mapped insertion or

FALSE in case it does not. Only columns with at least three mapped insertions were retained. This resulted in a Boolean matrix with 200 unique columns representing the insertion loci. To incorporate possible interactions with the genotype status of these tumors ($p19^{\mathrm{ARF}-/-}$, $p53^{-/-}$ or wt), we included three additional columns representing the three genotypes.

To measure association between the insertion loci (or combination of insertion loci) and the gene clusters we used a standard $t$-score. For each of the gene clusters we determined the single best locus with the strongest positive and negative association, as well as the best possible combination of loci for each of the 24 Boolean network topologies (Figure S1). Solutions with a permutation based $p$-value smaller than 0.001 were retained. In case multiple solutions for a single gene cluster remained, a rank aggregation approach (described in detail in the supplement), combining several measures of significance and biological relevance, was used to choose the most relevant model.

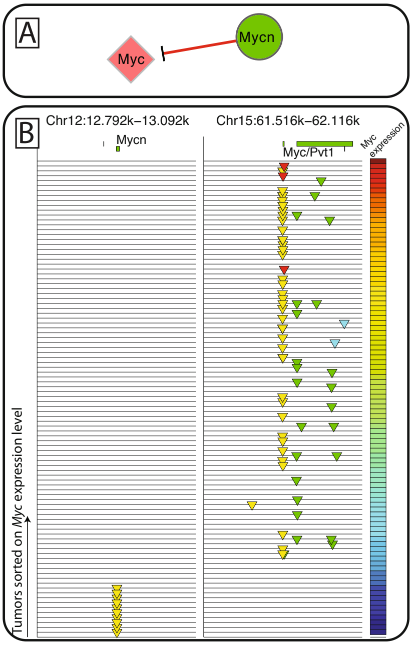
**Interaction Network.** Using this approach, we find significant ($p$ <0.001) single locus and mutilocus associations for 137 gene clusters (174 genes). A Cytoscape plot of these interactions is given in Figure S5. For 88 of the gene clusters, a single locus model, i.e. inserts at a single locus, was sufficient to obtain a significant segregation of the expression measurements. On the other hand, for 49 cases a more complex association was required to obtain a significant association (20 2-input networks and 29 3-input networks). Interestingly, the type of logic that was used in this set of significant interactions was depleted of AND logic. In fact, it was observed that AND logic generally showed poor association (irrespective of the $p$-value), suggesting that co-occurrence of insertions (i.e. insertions co-occurring in the same tumor, captured by AND logic) is not a common mechanism in regulating transcriptional activity.

**cis-eQTMLs.** Strong cis-associations, for which an insertion locus is associated with a proximal target transcript, are observed for insertions mapped to *Rras2*, *Ccnd1*, *Gfi1* and *Notch1*. In many other cases, direct association on the transcriptional level between insertions and their predicted targets is more subtle, i.e. the expression changes are very small, and fail to exceed the array noise. In other cases insertions may affect translation instead of transcription, and hence may not be detected in this analysis.

It is possible that the use of alternative routes of deregulating nearby genes dilutes the observed cis-association. This means that the absence of a mutation is no longer necessarily associated with low expression. A clear example of such a case is the expression of the *Myc* oncogene, which was found to be expressed (log2 expression level $> 7$) in 88 of the 97 tumors, while it harbored an insertion only in 51 tumors (Figure 3). This suggests that, in cases where an insertion near *Myc* is lacking, *Myc* is upregulated by other mechanisms. For most of the tumors in which *Myc* remains unexpressed, insertions near *Mycn* are observed. Indeed, our results reveal a strong negative association between insertions near *Mycn* and *Myc* expression (Figure 3). A plausible explanation for this observation is that *Mycn* insertions are functionally equivalent to insertions in the *Myc* locus, a mechanism which has been identified in human leukemias and lymphomas as well [19].

**Genotype Interactions.** By including three Boolean profiles representing the genotype we are able to retrieve genotype specific expression changes, as well as expression

**Fig. 3. Association between *Mycn* insertions and *Myc* expression. A)** The red diamond-shaped node represents the gene cluster containing, in this case, a single probe for *Myc*. The green circular node represents the insertion locus for *Mycn*. **B)** Locus plot of insertions in the *Mycn* locus and the *Myc* locus. Green (yellow) triangles denote positively (negatively) oriented oriented insertions that according to the literature rules were mapped to *Mycn/Myc*. Red (cyan) triangles denote positively (negatively) oriented insertions that were not mapped to a target gene. The color bar on the right represents expression levels of the *Myc* probe. Tumors were sorted based on the expression level of *Myc*.



changes that are due to putative interaction between genotype and one or two insertion loci (Figure 4). In addition to the probe for *p53* itself, many other well characterized targets of *p53* and *p19* were found among the direct associations identified by our analysis. More specifically, increase of *Cdkn2a* (*p19$^{ARF}$* isoform) expression is associated with the *p53*$^{-/-}$ tumors, suggesting a feedback loop mechanism compensating for the loss of *p53*. Interestingly, low expression of the *p16INK4a* isoform is found to be associated with wild-type tumors only, suggesting loss of the *p19/p53* pathway permits lymphoma development in the presence of increased p16 expression. Other known direct interactions include: *Bax* [20], *Cdkn1a* (p21) [21] and *Ccng1* (CyclinG1) [22] ) all of which are induced by p53. These examples demonstrate the robustness of our methodology.

A more complex association between genotype and transcript level was found in the case of *Usp18*, a gene which has been implicated in human non-small-cell lung cancer [23]. A 3-input network with the wild-type status, *p19*$^{\mathrm{ARF}-/-}$ status and the *Nfkb2/Sufu* locus was found to be negatively associated with low *Usp18* transcript levels. This network can be simplified to a 2-input OR network with *p53*$^{-/-}$ status and the *Nfkb2/Sufu* locus as inputs and a positive association with *Usp18* expression (Figure S4). Indeed, the *p53*$^{-/-}$ status was found to be strongly associated with elevated *Usp18* levels. However, in a substantial number of wild-type and *p19*$^{\mathrm{ARF}-/-}$ tumors elevated expression was also observed. Interestingly, the CAL network offers a partial explanation for this, since it reveals that three of the non-*p53*$^{-/-}$ tumors with high *Usp18* expression harbored insertions in the *Nfkb2/Sufu* locus. From this observation the interesting hypothesis can be derived that insertions near *Nfkb2/Sufu* offer an alternative to the loss of *p53* in upregulating *Usp18*.

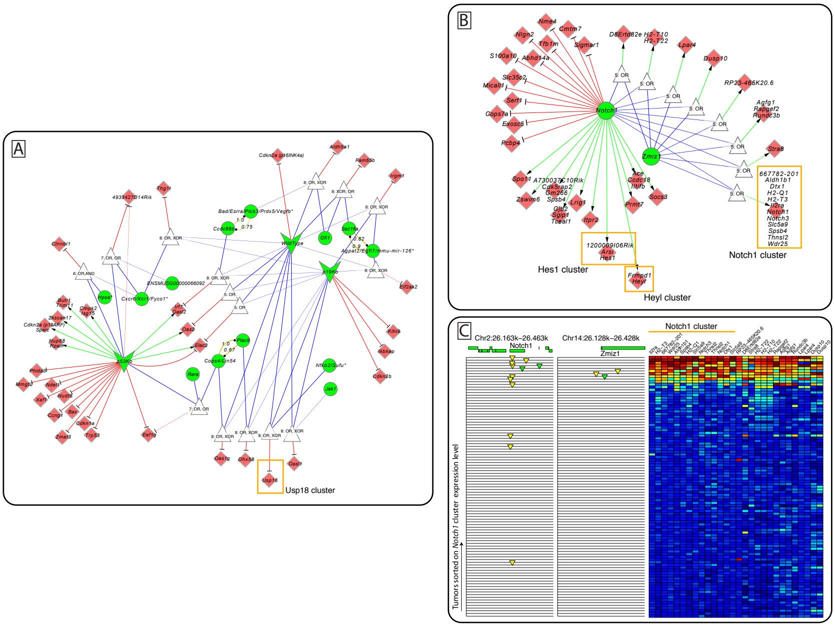


**Fig. 4. Cytoscape interaction diagrams of the interactions with the genotype (A) and *Notch1* B) status.** Green V-shaped nodes, green circular nodes, red diamond-shaped nodes represent the genotype status, insertion loci and gene clusters, respectively. The white triangles denote CAL networks, with the logic functions used specified in text. The number in the network nodes refer to the supplementary table. Green and red links represent positive and negative associations, respectively. The yellow links indicate proximal insertion loci, that share some of the mapped insertions. The numbers on these links indicate the fraction of insertions that are shared. In case the nodes are labeled with a (*), some genes were omitted from the complete list of putative targets for readability. Putative targets were only omitted in case literature revealed poor evidence for involvement in cancer or cell-functions like apoptosis or cell-cycle. A full list of putative targets is available in the online material (see Supplements for details). **C)** Locus plot of the *Notch1* and *Zmiz1* loci. For an explanation of the symbols see Figure 3. Only the probes at the output of a 2-input OR network with *Notch1* and *Zmiz1* are shown. Expression values were $z$-normalized to allow for comparison between probes.

**Regulatory Hubs.** The discovered interactions reveal that *Gfi1* and *Notch1* are clear hubs, and insertions in their vicinity are associated with expression of many transcripts. Interestingly, both genes have well established roles in cancer and moreover are known transcriptional regulators.

*Gfi1* encodes a nuclear zinc finger protein and is recognized to have different complex and cell context specific roles. In lymphoid cells, however, GFI1 is a known transcriptional repressor. This is consistent with the predominantly inhibitory interactions revealed by our analysis. The literature provides evidence for some of the putative regulatory interactions. An interesting example is negative association between inserts near *Gfi1* and transcript levels of *Btg1*. Human BTG1 is a known tumor suppressor and member of an anti-proliferative gene family that regulates cell growth and differentiation [24]. It has been implicated in acute lymphoblastic leukemia (ALL) [25] and non-hodgins lymphoma [26]. Association between *Gfi1* and *Btg1* activity may be explained as it was found that BTG1 is regulated by CEBPA [27], which, in turn, is a known target of GFI1 in T lymphocyte (Jurkat) cells [28].

Figure 4A shows the interaction diagram of associations of the *Notch1* locus and gene expression of multiple genes. In addition, the associations of a 2-input OR of *Notch1* and *Zmiz1* are shown. *Notch1* is a member of the family of NOTCH receptors, that operate both as recipients of extracellular signals at the cell surface and as transcription factors regulating gene expression in the nucleus. In its role as transcription factor, NOTCH1 forms a transcriptional activator complex and activates genes of the enhancer of split locus. Notably, *Hes1*, hairy and enhancer of split 1, and *Heyl*, a member of the hairy and enhancer of split-related (HESR) family, are both among the associated transcripts identified by our analysis. Both proteins have been implicated in cancer, and specifically implicated as targets of NOTCH signalling [29].

Using Chip-chip data previously published [30] of NOTCH1 and HES1 DNA binding in human T cell ALL cells [30], we checked if the orthologs of the *Notch1* target transcripts identified in our study were among the list of NOTCH1 bound genes. We found that 5 of the 23 *Notch1* targets with human orthologs were among the NOTCH bound target list (COPS7A, EXOSC5, HES1, ITPR2 and TFB1M). Since *Hes1* was among our *Notch1* targets, and it is possible that *Notch1* acts upon its targets through *Hes1*, binding of HES1 may explain the associations observed with *Notch1* mutations [31]. Therefore, we also checked for overlap of human orthologs of *Notch1* targets and the Chip-chip results of HES1 binding. In this way suggestive evidence for three additional interactions was found (CDK5RAP2, PRMT7 and TCEAL1).

**Multi-locus eQMLs Reveal Alternative Pathways.** Although *Notch1* insertions are found almost exclusively in tumors with elevated transcripts levels of *Notch1*, three tumors remain without *Notch1* insertions (Figure 4). One CAL network combines the *Notch1* locus with insertions in the *Zmiz1* locus. Insertions in the *Zmiz1* locus occur in tumors with elevated *Notch1* levels and two of these occur in tumors without insertions in the *Notch1* locus. Moreover, *Zmiz1* insertions are exclusively observed for tumors with elevated *Notch1*. A hypothesis worth exploring further is therefore that *Zmiz1* operates upstream of *Notch1* and, in case of the absence of a *Notch1* mutation, is able to upregulate *Notch1*.

## 3 Discussion

We propose mutational genomics, an approach to delineate transcriptional regulatory interaction networks in cancer by searching for associations in mutation data and gene expression measurements obtained from the same sample. When performed for a set of 97 lymphoid splenic tumors, an interaction network comprising 60 insertionally targeted loci and 174 putative target transcripts results. Because selective pressure exerted by the tumor growth enriches for loci with causal implications for tumorigenesis, many interactions in cancer related pathways were discovered.

A number of well characterized interactions were found, such as the association between loss of *p53* and reduced *Bax*, *Cdkn1a* and *Ccng1* levels. Known transcriptional regulators *Gfi1* and *Notch1*, both of which have established roles in tumorigenesis, were found to be associated to differential expression of many transcripts, suggesting a master regulator role for these genes in lymphomagenesis. The targets of insertions near *Notch1* included many genes whose promotors were found to be bound by NOTCH1 and or HES1 in human T cell ALL.

In addition to single locus associations, more complex associations were identified by inferring CAL networks, i.e. Boolean combinations of insertion loci. This revealed a possible role for insertions in the *Nfkb2/Sufu* locus in upregulating *Usp18* expression. Similarly, it was found that two of the tumors that did not appear to bear an activating *Notch1* mutation, harbored insertions in the *Zmiz1* locus, possibly explaining the elevated *Notch1* expression in these tumors. From this the hypothesis can be formulated that *Zmiz1* functions upstream of *Notch1*. This illustrates the potential of mutational genomics as a powerful way of generating hypotheses that can be validated in the lab.

While in this study we focused on retroviral insertional mutagenesis, transposon based insertional mutagenesis may be similarly suitable for mutational genomics [32]. This would greatly increase the number of tissues and tumor types in which mutational genomics can be employed, and thus increase the scope of this approach.

## 4 Materials and Methods

**Animal Experiments.** All animal experiments were done conform to national regulatory standards and are approved by the Animal Experiments Committee (DEC) of the Netherlands Cancer Institute (approval ID: OZP 02029).

**Gene Expression Preprocessing.** Gene expression measurements were obtained using the Illumina MouseWG6-V2 beadchips, and were normalized using VST and RSN. Probes without a map position were discarded. Only highly variant probes (within the top 25 percentile) were retained. Hierarchical clustering (complete linkage, correlation distance, distance threshold of 0.2) was employed to combined strongly correlated genes, resulting in 6261 clusters. A clipping filter was applied as described [17], to limit the effect of strong outliers, affecting 625 gene clusters. Finally, gene clusters for which the best possible split in two groups based on the $t$-score resulted in highly unbalanced class distribution (smallest class size of 3 or smaller), were removed. Altogether, this resulted in 6228 gene clusters that were used in the association analysis.

**Determining Insertion Loci.** The effect of insertions on the nearby targets is dependent on the relative position and orientation of the target transcript as well as the orientation of the viral integration [5, 7, 16]. To exploit this information, we have employed a rule-based mapping (RBM) procedure [33]. RBM associates each insertion to one or more putative target transcripts based on a set of rules that were distilled from literature (a more comprehensive description of RBM is given in the Supplements). The unique list of transcripts that follows from this procedure is used to generate binary profiles that, for each tumor, indicate if a transcript is a putative target. We observed that for proximal transcripts frequently the same binary profile results. These were therefore combined into a single profile. Insertion target profiles that contained transcript-insertion associations in more than three tumors were considered in the analysis and served as inputs for the association inference.

**CAL Network Inference.** CAL network inference has been described in detail [17]. Briefly, given some Boolean network topology $\mathcal{B}$, the objective is to find the combination of loci such that the association between the network output and some gene expression vector is optimal. Equivalently, we solve the following:

$$\underset{\mathbf{L}}{\operatorname{argmax}}\, f(\mathcal{B}(\mathbf{L}), \mathbf{g}), \tag{1}$$

where $\mathbf{L}$ is a $T \times N$ Boolean matrix containing $N$ input loci of length $T$, $\mathcal{B}$ is a Boolean function that maps the $N$ input loci to a single Boolean vector, $\mathbf{g}$ is a vector containing the expression values for some gene and $f$ is an association measure. Here, we use the $t$-statistic as the association measure. In the tail of the $t$-distribution an approximation of the $t$-score exists that can be optimized, using a branch-and-bound algorithm, in a fraction of the time required to optimize the real $t$-score.

To apply the CAL network inference approach to a dataset with $\sim$100 samples, some modifications to the original implementation of this method [17] were made in order to improve scalability further. All modifications are described in the Supplementary material.

**CAL Network Significance.** We solve Equation 1 for each gene cluster and for a range of 24 network topologies. The topologies are given in Figure S1. For each gene cluster-network topology combination a $p$-value can be obtained. The following procedure is performed to obtain the necessary null-distributions for each network topology separately. All 6228 gene clusters are permuted 90 times by shuffling the order of the clusters' gene expression values. This results in a total of 560k random permutations. For each permutation the CAL network search is performed, using the same parameter settings as were used on the real data. This results in 560k $t$-scores. The CAL network algorithm only produces reliable solutions above a certain tolerance level, which for these data was set to $t = 7.5$. We therefore calculate a piecewise cumulative distribution function (CDF). Below the tolerance level the CDF is set to zero, since in this region $t$-scores are not accurate. Above the tolerance, we use the empirical estimate of the CDF. A pseudocount is included to prevent $p$-values of zero.

In many cases it is possible to find strong (and significant) associations between the mutation data and gene expression using several network topologies. In order to select the most biologically relevant model, we rank all solutions based on several other measures of significance and biological relevance. These measures include: 1) the $p$-value improvement compared to the lowest $p$-value obtained for networks with fewer inputs, 2) the number of inputs of the network topology, 3) the coverage of the truth table of the network topology, 4) the number of samples in the smallest class. Average Borda ranking is used to aggregate ranks from these four measures [34]. Only solutions that receive the highest rank are reported.

**Acknowledgements.** This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). This research has partially been supported by the Dutch Life Science Grid initiative of NBIC and the Dutch e-Science Grid BiG Grid, SARA - High Performance Computing and Visualisation.

## References

1. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. Cell 144, 646–674 (2011)
2. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530–536 (2002)
3. van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A.M., et al.: A gene-expression signature as a predictor of survival in breast cancer. N Engl. J. Med. 347, 1999–2009 (2002)
4. Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., et al.: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. USA 98, 10869–10874 (2001)
5. Kool, J., Berns, A.: High-throughput insertional mutagenesis screens in mice to identify oncogenic networks. Nature Reviews Cancer 9, 389–399 (2009)
6. Kool, J., Uren, A.G., Martins, C.P., Sie, D., de Ridder, J., et al.: Insertional mutagenesis in mice deficient for p15ink4b, p16ink4a, p21cip1, and p27kip1 reveals cancer gene interactions and correlations with tumor phenotypes. Cancer Res. 70, 520–531 (2010)
7. Uren, A.G., et al.: Retroviral insertional mutagenesis: past, present and future. Oncogene 24, 7656–7672 (2005)
8. Mikkers, H., Berns, A.: Retroviral insertional mutagenesis: tagging cancer pathways. Adv. Cancer Res. 88, 53–99 (2003)
9. Jansen, R.C., Nap, J.P.: Genetical genomics: the added value from segregation. Trends Genet. 17, 388–391 (2001)
10. Gerrits, A., Dykstra, B., Otten, M., Bystrykh, L., de Haan, G.: Combining transcriptional profiling and genetic linkage analysis to uncover gene networks operating in hematopoietic stem cells and their progeny. Immunogenetics 60, 411–422 (2008)
11. Li, J., Burmeister, M.: Genetical genomics: combining genetics with gene expression analysis. Hum. Mol. Genet. 14(spec. 2), R163–R169 (2005)
12. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., et al.: Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302 (2003)
13. Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., et al.: Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat. Genet. 37, 225–232 (2005)
14. Gerrits, A., Li, Y., Tesson, B.M., Bystrykh, L.V., Weersing, E., et al.: Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet 5, e1000692 (2009)
15. Erkeland, S.J., Verhaak, R.G.W., Valk, P.J.M., Delwel, R., Löwenberg, B., et al.: Significance of murine retroviral mutagenesis for identification of disease genes in human acute myeloid leukemia. Cancer Res. 66, 622–626 (2006)
16. Jonkers, J., Berns, A.: Retroviral insertional mutagenesis as a strategy to identify cancer genes. Biochim. Biophys. Acta 1287, 29–57 (1996)
17. de Ridder, J., Gerrits, A., Bot, J., de Haan, G., Reinders, M., et al.: Inferring combinatorial association logic networks in multimodal genome-wide screens. Bioinformatics 26, i149–157 (2010)
18. Uren, A.G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., et al.: Large-scale mutagenesis in p19(arf)- and p53-deficient mice identifies cancer genes and their collaborative networks. Cell 133, 727–741 (2008)
19. Hirvonen, H., Hukkanen, V., Salmi, T.T., Pelliniemi, T.T., Alitalo, R.: L-myc and n-myc in hematopoietic malignancies. Leuk Lymphoma 11, 197–205 (1993)
20. Chipuk, J.E., Kuwana, T., Bouchier-Hayes, L., Droin, N.M., Newmeyer, D.D., et al.: Direct activation of bax by p53 mediates mitochondrial membrane permeabilization and apoptosis. Science 303, 1010–1014 (2004)

21. Dulić, V., Kaufmann, W.K., Wilson, S.J., Tlsty, T.D., Lees, E., et al.: p53-dependent inhibition of cyclin-dependent kinase activities in human fibroblasts during radiation-induced g1 arrest. Cell 76, 1013–1023 (1994)
22. Komarova, E.A., Diatchenko, L., Rokhlin, O.W., Hill, J.E., Wang, Z.J., et al.: Stress-induced secretion of growth inhibitors: a novel tumor suppressor function of p53. Oncogene 17, 1089–1096 (1998)
23. Lam, D.C.L., Girard, L., Ramirez, R., Chau, W.S., Suen, W.S., et al.: Expression of nicotinic acetylcholine receptor subunit genes in non-small-cell lung cancer reveals differences between smokers and nonsmokers. Cancer Res 67, 4638–4647 (2007)
24. Rouault, J.P., Rimokh, R., Tessa, C., Paranhos, G., Ffrench, M., et al.: Btg1, a member of a new family of antiproliferative genes. EMBO J. 11, 1663–1670 (1992)
25. van Galen, J.C., Kuiper, R.P., van Emst, L., Levers, M., Tijchon, E., et al.: Btg1 regulates glucocorticoid receptor autoinduction in acute lymphoblastic leukemia. Blood 115, 4810–4819 (2010)
26. Morin, R.D., Mendez-Lago, M., Mungall, A.J., Goya, R., Mungall, K.L., et al.: Frequent mutation of histone-modifying genes in non-hodgkin lymphoma. Nature 476, 298–303 (2011)
27. Tavor, S., Park, D.J., Gery, S., Vuong, P.T., Gombart, A.F., et al.: Restoration of c/ebpalpha expression in a bcr-abl+ cell line induces terminal granulocytic differentiation. J. Biol. Chem. 278, 52651–52659 (2003)
28. Duan, Z., Horwitz, M.: Targets of the transcriptional repressor oncoprotein gfi-1. Proc. Natl. Acad. Sci. U S A 100, 5932–5937 (2003)
29. Katoh, M., Katoh, M.: Integrative genomic analyses on hes/hey family: Notch-independent hes1, hes3 transcription in undifferentiated es cells, and notch-dependent hes1, hes5, hey1, hey2, heyl transcription in fetal tissues, adult tissues, or cancer. Int. J. Oncol. 31, 461–466 (2007)
30. Margolin, A.A., Palomero, T., Sumazin, P., Califano, A., Ferrando, A.A., et al.: Chip-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc. Natl. Acad. Sci. U S A 106, 244–249 (2009)
31. Dudley, D.D., Wang, H.C., Sun, X.H.: Hes1 potentiates t cell lymphomagenesis by up-regulating a subset of notch target genes. PLoS One 4, e6678 (2009)
32. Mattison, J., van der Weyden, L., Hubbard, T., Adams, D.J.: Cancer gene discovery in mouse and man. Biochim. Biophys. Acta 1796, 140–161 (2009)
33. de Jong, J., de Ridder, J., van der Weyden, L., Sun, N., van Uitert, M., et al.: Computational identification of insertional mutagenesis targets for cancer gene discovery. Nucleic Acids Res 39, e105 (2011)
34. Lin, S.: Rank aggregation methods. Wiley Interdisciplinary Reviews: Computational Statistics (2010)