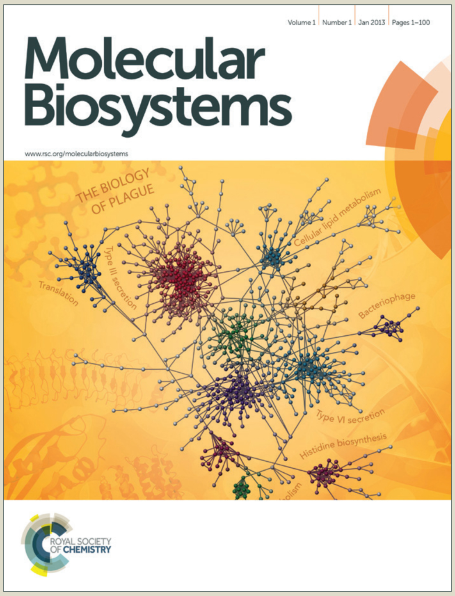

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

Mutated Genes and Driver Pathways Involved in Myelodysplastic Syndromes—A Transcriptome Sequencing Based Approach

Liang Liu[1*], Hongyan Wang[1*], Jianguo Wen[2*], Chih-En Tseng[2,3*], Youli Zu[2], Chung-che Chang[4§], Xiaobo Zhou[1§]

[1] Center for Bioinformatics and Systems Biology, Division of Radiologic Sciences, Wake Forest University Baptist Medical Center, Winston-Salem, NC 27157, USA.

[2] Department of Pathology, the Methodist Hospital Research Institute, Houston, TX 77030, USA.

[3] Buddhist Dalin Tzu Chi Hospital; School of Medicine, Tzu Chi University, Hualien, Taiwan

[4] Department of Pathology, University of Central Florida, Orlando, FL 32803, USA

*These authors contributed equally to this work

§Corresponding author

Email addresses:

LL: lliu@wakehealth.edu

HW: wanghongyan.pku@gmail.com

JW: jwen@tmhs.org

CET: p121521@gmail.com

YZ: yzu@tmhs.org

CCC: C.Jeff.Chang.MD@Flhosp.org

XZ: xizhou@wakehealth.edu

Abstract

Background

Myelodysplastic syndromes are a heterogeneous group of clonal disorders of hematopoietic progenitors and have potentiality to progress into acute myelogenous leukemia. Development of effective treatments has been impeded by limited insight into pathogenic pathways. In this study, we applied RNA-seq technology to study the transcriptome on 20 MDS patients and 5 age-matched controls, and developed a pipeline for analyzing this data. After analysis, we identified 38 mutated genes contributing to MDS pathogenesis. 37 out of 38 genes have not been reported previously, suggesting our pipeline is critical for identifying novel mutated genes in MDS. The most recurrent mutation happened in gene IFRD1, which involved 30 % of patient samples. Biological relationships among these mutated genes were mined using Ingenuity Pathway Analysis, and the results demonstrated that top two networks with highest scores were highly associated with cancer and hematological diseases, indicating that the mutated genes identified by our method were highly relevant to MDS. We then integrated the pathways in KEGG database and the identified mutated genes using our novel rule-based mutated driver pathway scoring approach for detecting mutated driver pathways. The results indicated two mutated driver pathways are important for the pathogenesis of MDS: pathway in cancer and in regulation of actin cytoskeleton. The latter, which likely contributes to the hallmark morphologic dysplasia observed in MDS, has not been reported, to the best of our knowledge. These results provide us new insights into the pathogenesis of MDS, which, in turn, may lead to novel therapeutics for this disease.

Keywords: Myelodysplastic Syndromes, Mutations, Mutated Driver Pathways

Introduction

The frequency and incidence of myelodysplastic syndromes (MDS) are increasing in the US population. This is attributed to multiple factors, including advancing age, the use of cytotoxic and mutagenic therapies for cancer, and greater exposure to environmental toxins [1]. We have recently reviewed the pathobiology and molecular pathogenesis of MDS and concluded that the etiology and pathogenesis of MDS are very complex and remain inadequately characterized [2]. MDS is considered a clonal stem cell disease characterized by peripheral cytopenia with normocellular or hypercellular marrow and monolineage, bilineage, or trilineage dysplasia [3]. According to the WHO classification, the main categories of MDS (listed in order of increasing severity) include refractory anemia (RA), RA with ringed sideroblasts (RARS), refractory cytopenia with multi-lineage dysplasia (RCMD), and RA with excess blasts (RAEB) [4]. Currently, a comprehensive understanding of the pathogenesis remains largely elusive, and thus outcomes in MDS patients have not improved over the last few decades [5].

The completion of the Human Genome Project has enabled molecular genetic classification of lymphoid and myeloid malignancies using high-throughput genome-wide approaches (e.g., cDNA microarrays, array CGH and SNP arrays) [6]. These approaches, applied to studies of MDS since the last decade, have resulted in conflicting and inconsistent results, regarding to how to differentiate low-risk from high-risk MDS patients [7]. In addition, these results cannot pinpoint genes responsibility, or how they cause MDS phenotypes. Answers to these questions are needed to elucidate the molecular mechanisms behind MDS, and to influence the design and development of new strategies for its diagnosis, treatment, and prevention.

Next generation sequencing (NGS) is a new methodology for whole genome or transcriptome research, and enables cost-effective sequencing with higher sensitivity. In contrast to microarray-based studies, an NGS-based approach can generate single base-resolution data, and enable high-resolution identification of point mutations, small insertions/deletions, alternatively spliced transcripts, alternatively polyadenylated transcripts, and fusions that may be common in cancers or MDS [8].

Some putative mechanism-related mutations in MDS have been identified, including classical oncogenes, such as *RUNX1* and *TP53,* which were associated with severe thrombocytopenia and an increased proportion of bone marrow blasts [9]. Recent studies also found some recurrent mutations on splicing factors in MDS [10-12]. These mutated genes, including *U2AF35*, *ZRSR2*, *SRSF2*, *SF3B1*, *SF3A1*, *SF1*, *U2AF65*, and *PRPF40B*, are involved in multiple components of the RNA splicing machinery.

Many methods have been proposed to distinguish the mutations called "driver mutations" (promoting cancer cell proliferation) from those called "passenger mutations" (being neutral to cancer proliferation). A standard approach for recognizing a "driver" mutation is to detect recurrently mutated genes in cancer samples. However, this is not an effective way to reveal all driver mutations, since different mutations may target the same pathway and genomic aberrations in cancer exhibit significant mutational heterogeneity, even among the same histologic type of cancers [13, 14]. As a result, some researchers have shifted to a focus on mutated driver pathways, and the identification of cancer pathways [15-17].

Recent studies uncovered two features associated with mutated driver pathways [18]. First, mutated driver pathways were found in many patients, which demonstrate characteristics of "high coverage"; second, mutated driver pathways obey the "exclusivity principle", which means most patients only have one mutation in a driver pathway. Thus, some measurements and algorithms have been proposed to recognize mutated driver pathways [19, 20]. However, these methods have generally assumed that mutated pathways have similar gene expression, so they may not work for cancers that are more heterogeneous in nature, such as MDS.

The purpose of this study was to explore novel MDS-related mutated genes by developing a novel bioinformatical pipeline. In samples from 20 MDS patients and 5 age-matched controls, we selected the mutations, and predicted to be deleterious to the function of corresponding genes. We then discovered mutated driver pathways, using a unique method, named a rule-based mutated driver pathway scoring approach (RMDP^{score}), to integrate mutation information with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, so as to identify the mutated driver pathways that may contribute to the pathogenesis of MDS. Our findings may lead to the development of novel therapeutics for this disease.

Results

Mutated genes

The NGS-based methodology for mining MDS-related mutations is summarized in **Figure** 1. First, RNA was extracted from HSCs and analyzed using Genome Analyzer II from Illumina. Then, tools were adopted to align reads to whole human genome and call variants. After calling variants, removing known single nucleotide polymorphisms (SNPs) and deleting synonymous mutations, the 20 MDS samples revealed 91 mutated genes when we set the threshold of coverage as 30. Six samples had no mutations. Most of the mutated genes were mutated in one sample; 12 genes mutated in two or more samples.

Mutations may change the sequence of a gene without affecting its function. To find function-involved mutations, we used PROVEAN [21] and SIFT [22] to assess the biological effect of these mutations. We found 38 mutated genes that were non-synonymous, involving 44 mutated points (see **Table 1** for details). All these genes mutated in one sample except for the *IFRD1*.

The most recurrent mutated gene was C133R mutation interferon-related developmental regulator 1 gene (*IFRD1*), which occurred in 6 samples (30%). *FRD1* could play a role in regulating gene activity in the proliferation and differentiation pathways induced by nerve growth factor (NGF). In humans, a non-synonymous variant in *IFRD1* is verified as a disease-causing candidate [23]. Although *IFRD1* mutations have been reported previously in sensory/motor neuropathy with ataxia [23] and cystic fibrosis lung disease [24], they have not previously been associated with MDS. Among mutations listed in Table 1, only one gene, *IDH1*, was reported previously as associated with MDS [9]. Of note, although some results obtained from PROVEAN and SIFT may disagree, they consistently predicted 26 mutations, including 2 nonsense ones.

To verify the mutation detection pipeline, we chose 3 genes (*IFRD1*, *IQGAP2*, and *DIDO1*) for validation (See **Mutation validation** section for details). *IFRD1* locates in 7q31.1, *IQGAP2* locates in 5q13.3, and *DIDO1* locates in 20q13.33. 5q, 7q and 20q are all located in frequent deletion segments of the chromosomes in MDS patients. The chromosomal deletion may result in haploinsufficiency of genetic expression and/or the mutation-related loss of genetic functions possibly committed to pathogenesis of MDS. We used chromatography (**Figure 2**) to verify the presence of these 3 mutations in our samples.

Although single nucleotide mutations cannot reveal how they affect the pathway directly, they do affect the function of corresponding gene. According to CanPredict [25] (http://www.canpredict.org/) the point mutation of IQGAP2 with glutamine to histidine substitution (Q1146H) results in likely cancerous changes, which indicates that loss of its tumor suppressor function might contribute to the pathogenesis of MDS. We also studied the structure of *IQGAP2* and found that the mutated nucleotide was located in its GTPase domain, which may impair its GTPase activity (**Figure 3**).

IPA analysis

We further merged the detected 38 mutated genes with 86 mutated genes previously reported in the literature for IPA analysis. Network analysis tools in IPA were used to determine the biological relationships among MDS-related candidates. The two networks with the highest scores were highly associated with cancer and hematological disease (**Figure 4**). Of note, IPA analysis was also performed to the gene set consisting of only 38 genes detected in this work. The cancer-related network centered at CDKN2A and NFκB genes were still highly enriched; however the one associated with hematological disease was not observed. The first network includes 26 genes (gray and red nodes) we input; 11 (red nodes) came from our pipeline. This network centers at CDKN2A and NFκB, and the mutated genes are very close to them. Protein ROCK2, ERAP2, and NXF1 directly connect with NFκB. ROCK2 is one isoform of the ROCK genes, which are important regulators of cell growth, migration, and apoptosis via control of actin cytoskeletal assembly. The mutation we detected, Y285H, locates in the kinase domain of ROCK2 (**Figure 3**), which may affect its phosphorylation function [26]. The relationship between ROCK2 and NFκB demonstrates that inhibiting active ROCK2 will decrease activation of the NFκB complex, indicating that functional mutation of ROCK2 may affect the function of NFκB. The other target of ROCK2 is 26s proteasome, which is the ATP-dependent proteolytics complex, responsible for ubiquitin-dependent protein degradation. The IPA results indicate that Rock2 protein mediates activation of the 26s proteasome in Neuro-2a cells of mouse. Similarly, NXF1 protein (TAP) contributes to the activity of NFκB complexes according, and its mutation may also decrease NFκB activity. NFκB plays a major role in hematopoiesis, and is important in the development of hematopoietic lineages. Recent data strongly show that the deregulation of NFκB pathway significantly contributes to many hematopoietic cell diseases [27].

Another network (**Figure 4B**) only contains three genes we detected (red nodes). However, the most recurrently mutated gene in our sample, *IFRD1*, is involved. *IFRD* is involved in myocyte differentiation, and it regulates the NFκB pathway by affecting the activity of histone deacetylases and forming complexes with the p65 subunit of NFκB [28].

Mutated driver pathway analysis

To comprehensively identify mutated driver pathways, we used a rule-based mutated driver pathway scoring approach (RMDP^{scoring}) (see Methods) with an adjusted threshold of coverage to 15 when calling mutations (**Figure 1**). With this approach, 470 mutated genes were identified, including 13 genes (*GNAS*, *IDH1*, *PTPN11*, *RUNX1*, *CTSL1*, *U2AF1*, *XPO1, CSTF1*, *CTSL1*, *HNRNPCL1*, *NAA16*, *SDCCAG1*, and *SF3B1*) previously reported as associated with MDS. Among these 470 genes, 129 genes were annotated in the KEGG pathway database, and were used for the mutated driver pathway analysis. The mutation matrix and connection matrix were constructed as described in the Methods section.

We believe the chance that many mutated genes enrich in a common pathway is small. Therefore, k was selected in the range between 4 and 10. For each k , our algorithm outputted several gene sets with highest score. Then from the outputs, we further selected gene sets that perfectly match the definition of the mutated driver pathway. When $k = 6$, we obtained two optimal gene sets involved in cancer pathways (**Figure 5A**, **Figure 5B** and **Figure 6**). These two sets were confined to only 8 different genes: *FLT3*, *HSP90B1*, *MAPK9*, *MLH1*, *TPM3*, *TPR*, *SMAD4*, and *RUNX1*. The fms-related tyrosine kinase 3 gene (*FLT3*) encodes tyrosine-protein kinase, acting as a cell-surface receptor for the cytokine FLT3LG. The FLT3 pathway is one of the most important pathways in blood disease, because it regulates differentiation, proliferation, and survival of hematopoietic progenitor cells. Previous studies reported that *FLT3* mutation lead to constitutive kinase activation; its mutations represent most recurrent perturbations in acute myeloid leukemia (AML) [29, 30]. Our sequencing data identified the mutation L832W in *FLT3*. This mutation locates in the kinase domain, like most mutations of *FLT3* found in AML [30], and was predicted to be deleterious by PROVEAN and to be damaging by SIFT. The *HSP90B1* (G543R) mutation in our samples locates in the ribosomal protein S5 domain. This is also a harmful mutation, according to the PROVEAN and SIFT results. *MAPK9* encodes a member of the MAP kinase family, which is involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development. We found that it has a mutation (F225S) in the protein kinase domain and this mutation also damages its function.

Aside from the known downstream pathways of *FLT3* (such as PI3K, Ras, STAT5, and RUNX1), SMAD4, and MLH1 may also be regulated by FLT3 [31]. They take parts in anti-apoptosis, abnormal cell growth, and blocking differentiation. Hsp90 is also involved in this regulation and following abnormal hematopoiesis by chaperoning FLT3 and downstream signaling molecules, as well as direct effects on mitochondria [32]. The functional profiles of these genes are summarized in **Figure 6**.

When $k = 4$, we found 2 optimal gene sets (*ABI2*, *IQGAP2*, *PPP1R12A*, *ROCK1*) and (*ABI2*, *IQGAP2*, *PPP1R12A*, *PIP4K2B*), five genes in total. These are involved in regulation of actin cytoskeleton (**Figure 5C** and **Figure 5D**). The Abi-2 (protein of *ABI2*) may regulate cell growth and transformation by interacting with non-receptor tyrosine kinases. However, the function of Abi-2 still remains largely unknown.

Discussion

MDS includes a series of hematopoietic stem cell disorders with high heterogeneity. This characteristic limits the efficiency of a research approach based only on mutation frequency of particular genes [10-12]. Instead of recurrent mutations, which were the focus of prior studies [10-12], our pipeline aims to recognize functional mutations and mutated driver pathways. First, we used function analysis tools, such as PROVEAN and SIFT, to estimate the effect of mutations. Our method found a promising marker, called *IFRD1*, which was mutated in 30 % of samples. This gene has a major role in sensory/motor neuropathy with ataxia and cystic fibrosis [23, 24]. Two other genes of interest are *IQGAP2* and *DIDO1*. Like *IFRD1*, they all located in frequent deletion segments of the chromosomes in MDS patients. *IQGAP2* is a putative tumor suppressor; down-regulation and inactivation through aberrant promoter methylation of *IQGAP2* were found in hepatocellular carcinoma and promote invasion in gastric cancer, respectively [33]. The *DIDO1* (death inducer-obliterator 1) gene located on 20q13.33, is another putative tumor suppressor gene. It is up-regulated early in apoptosis and implicated in the induction of myeloproliferative neoplasm/myelodysplastic syndrome [34]. PROVEAN and SIFT both predicted that mutations in *IQGAP2* and *DIDO1* were damaging. Additionally, structure analysis indicates that some mutations happen in important motifs; for example, the *IQGAP2* mutation is on its GAPase domain. These novel discoveries are significant given their major roles in various diseases and cancers [33, 35], suggesting that these genes may also contribute to the development of MDS.

IPA analysis indicated that the mutated and transcribed genes identified were concentrated in two networks that are highly associated with cancers and hematological diseases. These results demonstrated the importance of detected mutated genes to MDS. This also reduces one shortcoming of the current study – inability of validating MDS-free somatic tissue in the same patient, to rule out the possibility of germ line mutation/polymorphism. Although IPA analysis can discern networks among mutated genes, not all genomic aberrations in malignant cells contribute to the malfunction of cell growth, proliferation, and apoptosis.

We then used our unique RMDP^{scoring} methodology to identify the mutated driver pathways, i.e. a set of mutated genes that lead to development of MDS. This approach indicates that two key pathways (i.e. cancer and regulation of actin cytoskeleton) may play important roles in MDS. The pathway in cancer is likely related to the pre-leukemic status of MDS, and the other may contribute to the hallmark morphologic dysplasia observed in MDS due to the dysregulation of the cytoskeleton.

Our approach is particularly important for a disease that is very heterogeneous in nature, such as MDS. According to previous studies, recurrent mutated genes occurred at very low frequencies in MDS patients. This heterogeneity in mutated genes is reinforced by the fact that 37 of the 38 mutated genes we found to be associated with MDS were not reported by previous studies when using genomic sequencing approaches. This indicates that our RNA-seq approach is likely more efficient for identifying novel mutated genes in MDS, since this approach focuses on genes that are transcribed and thus functional.

Compared with previous driver mutated pathway analysis methods [19], the advantage of RMDP^{scoring} methodology is that it integrates well-studied KEGG pathways instead of using pure gene expression data. This idea was motivated by two concerns. First, there is no strong evidence to support that co-expressed genes in MDS had interactions, and expression data have a low signal-to-noise ratio. Second, MDS is highly heterogeneous, and relationships between genes will not be consistent in different patients. However, our method is somewhat limited by using prior well-established pathways; these are expanding dramatically with the new undergoing "omics" studies. Thus, the strength of RMDP^{scoring} as a new enrichment analysis which concerns the characteristic of "driver pathways" will be further improved.

Our work demonstrated the important roles of gene mutations in MDS, and delivered new insights to the pathogenesis. Limitations exist in the currently available data samples, which may prevent us from other important discoveries. For example, with the limited size of patient samples, correlations between mutations and severity/gender of MDS patients cannot be investigated at this time. In future, we will continue to collect samples, and expect to provide more comprehensive studies.

Materials and Methods

This study was approved by the Institutional Review Board of The Methodist Hospital, Houston, Texas, USA and the need for written informed consent from the participants was waived by the IRB.

Primary samples and NGS

RNA samples were prepared for next generation sequencing from 25 cases (http://ctsb.is.wfubmc.edu/MDS/MDS.html), including 5 age-matched controls and 20 MDS patients. Individuals ranged from 30 to 78 years old (mean age 58.6 years). Among those with MDS, we had 3 cases of refractory cytopenia with multi-lineage dysplasia (RCMD); 7 cases of RA with excess blasts (RAEB); 1 case of acute myelogenous leukemia with MDS; 9 cases of MDS with subtype unknown. First, cryopreserved bone marrow aspirates were thawed. Then, hematopoietic stem cells (HSCs) were sorted out from each bone marrow aspirate by CD34 magnetic beads (Miltenyi Biotec, Auburn, CA). The HSCs were directly lysed in Nugen lysis buffer (Nugen, San Carlos, CA), then transcribed into cDNA with the WT-Ovation RNA Amplification System (Nugen) and analyzed according to the user's guide. After sonication, the fragmented cDNA from 200 to 400 base pairs was isolated and amplified with 10 cycles of a polymerase chain reaction assay according to the paired-end protocol for the Genome Analyzer II (GAII) (Illumina, San Diego, CA). The prepared cDNA library was then sequenced on the GAII [36]. The reads (short cDNA sequences) obtained from the GAII were mapped to the reference human transcript database (NCBI, National Center for Biotechnology Information) by the methods described below.

NGS data analysis

We developed a pipeline (**Figure 1**) to detect functional mutated genes and mutated driver pathways from the MDS RNA-seq data. Following next generation sequencing, raw reads were first aligned to the UCSC reference genome using Bowtie [37]. Then the outputs were further processed by SAMTOOLS [38] for piling up the alignments and Seqgene [39] for mutation calling and annotation. After this step, we excluded known SNPs, mutations in no-coding regions, mutations that also happened in controls and mutations that were synonymous. The synonymous mutations were removed due to their unclear functions and the availability of functional annotation tools. For example, the two tools applied in this work, PROVEAN [21] and SIFT [40], can only be used for function annotation of non-synonymous mutations. The mutated genes that remained were candidates for further analysis. In our pipeline, we set two levels of threshold for the coverage of called variants. One was 30, which ensured low false positive rates. Mutations with coverages higher than 30 were selected for functional analysis. Meanwhile, for the mutated driver pathway study, we used a threshold of 15, allowing for a more comprehensive mutated driver pathway analysis.

Biological Function of Mutations

Not all mutations damage the function of corresponding genes. Some tools, for example PROVEAN[41] (Protein Variation Effect Analyzer) and SIFT[22] (Sorting Intolerant From Tolerant) (both available at http://www.jcvi.org/cms/research/software/), can identify non-synonymous variants that should be functionally important. We input all mutations detected into these two tools to predict if these mutations were deleterious (damaging) or neutral (tolerated). Mutations that were predicted to be either deleterious by PROVEAN or damaging by SIFT were selected for downstream analysis.

Mutation validation

To validate the reliability of our mutation detection method, the mutations in 3 genes (*IFRD1*, *IQGAP2*, and *DIDO1*) were further validated by Sanger sequencing. These genes were amplified from the aliquot of cDNA library for NGS with following primers: 5'-TGCGAAGACAAGGCAAGCAGC and 5'-TCAATTCCAGGGCCCAGCTGA for *IFRD1*, 5'-TCGACATGACAGCTGGAGGTCAGA and 5'-AACTTCTCTTCTGGCTCAGGGACA for *IQGAP2*, 5'-GCCTGTGATAGAGGCGCCTGG and 5'-GCCTGAGCACCGCATTTCGG for *DIDO1*. PCR products were purified using QIAquick PCR Purification Kit (Qiagen, Valencia, CA), and then sequenced using Sanger method at Baylor College of Medicine Sequencing Core Laboratory.

Ingenuity Pathway Analysis (IPA)

The IPA database (http://www.ingenuity.com/products/pathways_analysis.html) consists of millions of interactions between proteins extracted from the biological literature; thus, it can be used to indicate the likelihood of the input genes or proteins in a given pathway or network. We collected 114 genes and input them into the IPA database for network analysis. Of the 114 genes, 38 were recognized by our mutation detection pipeline (**Figure 1**). The other 86 genes which may be mutated in some cases of MDS were retrieved from three published studies [10-12] .

Mutated driver pathway analysis

We first formalized the mutated driver pathway discovery method by following Vandin's work [20]. Then we proposed a rule-based mutated driver pathway scoring approach (RMDP^{score}) by integrating the formalization with validated pathways in the KEGG database (downloaded from the Broad Institute).

In [20], the characteristics of the mutated driver pathway were abstracted as a Maximum Weight Sub-matrix problem:

$$\arg\max_{M} f(M) = |\Gamma(M)| - \omega(M) \qquad (1)$$

where $A_{m\times n}$ is mutation matrix, m is number of samples and n is the number of genes, $M_{m\times k}$ is sub-matrix of $A_{m\times n}$ and k is the number of genes selected in mutated driver pathways. $\Gamma(M) = \bigcup_{g\in M} \Gamma(g) = \bigcup_{g\in M} \{i : A_{ig} = 1\}$ is a set of patients who have mutations in a gene subset corresponding to matrix M ; $\omega(M) = \sum_{g\in M} \Gamma(g) - |\Gamma(M)|$ measures the coverage of M , which means $-\omega(M)$ can be viewed as a measurement of exclusivity.

Based on this formalization, Vandin *et al.* and Zhao *et al.* separately proposed Markov Chain Monte Carlo (MCMC) [19, 20]and Binary Linear Programming (BLP) [19]methods to optimize function. Zhao *et al*. [19]also considered biological context by integrating gene expression profile. However, gene expression has a low signal-to-noise ratio that may distort the interactions between genes. In this research, we were interested in the mutated driver pathways related to MDS. Therefore, instead of gene expression data, we integrated validated pathways (KEGG) into the above model to get a rule-based mutated driver pathway scoring approach:

$$\arg\max_{M} f(M) + \lambda * C(N) \ , \qquad (2)$$

where $N_{k\times k}$ is a sub-matrix of B with $b_{ij} = 1$ if gene i and gene j are in the same pathway and $b_{ij} = 0$ otherwise, it corresponds to the same gene set as in the sub-matrix M ; $C(N) = \sum_{i>j} 2 * n_{ij} / (k * (k-1))$ measures the connectivity of these genes included in matrix N . λ is a parameter set as $2 * k$ in our case. The scoring function in (2) was solved using Genetic algorithm (GA) [42, 43]. For every pre-determined k , GA was run ten times to eliminate the effect caused by different initial value settings.

Conclusion

In this study, we demonstrated the feasibility of using a new NGS-based transcriptomic methodology to study the pathogenesis of MDS. Using the bioinformatics pipeline established herein, this study identified key MDS-related mutated genes, and constructed potential driver pathways with a rule-based score approach RMDP$^{\text{scoring}}$. This knowledge will provide further insights into MDS, highlight the role of mutated genes and driver pathways, and accelerate the development of new therapeutic agents.

List of abbreviations

MDS : Myelodysplastic syndromes

RMDP$^{\text{score}}$: rule-based mutated driver pathway scoring approach

RA : refractory anemia

RAEB : RA with excess blasts

IPA : Ingenuity Pathway Analysis

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to acknowledge the members of Center for Bioinformatics and Systems Biology at Wake Forest University School of Medicine. This work was supported by National Library of Medicine [5R01LM010185-04] (Zhou).

References

1. Kasper DLM, Fauci ASM, Longo DLM, Braunwald EM, Hauser SLM, Jameson JLM, PhD (eds.): **Harrison's Principles of Internal Medicine - 16th Ed. (2005)**, 16th edn. Printed in the United States of America: The McGraw-Hill Companies, Inc.; 2005.
2. Nishino HT, Chang CC: **Myelodysplastic syndromes: clinicopathologic features, pathobiology, and molecular pathogenesis**. *Arch Pathol Lab Med* 2005, **129**(10):1299-1310.
3. Hofmann WK, Koeffler HP: **Myelodysplastic syndrome**. *Annu Rev Med* 2005, **56**:1-16.
4. Harris NL, Jaffe ES, Diebold J, Flandrin G, Muller-Hermelink HK, Vardiman J, Lister TA, Bloomfield CD: **The World Health Organization classification of neoplasms of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting--Airlie House, Virginia, November, 1997**. *The hematology journal : the official journal of the European Haematology Association / EHA* 2000, **1**(1):53-66.
5. Liu Y, Asai T, Nimer SD: **Myelodysplasia: battle in the bone marrow**. *Nature medicine* 2010, **16**(1):30-32.
6. Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF, Tibshirani R, Dohner H, Pollack JR: **Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia**. *N Engl J Med* 2004, **350**(16):1605-1616.
7. Pellagatti A, Cazzola M, Giagounidis AA, Malcovati L, Porta MG, Killick S, Campbell LJ, Wang L, Langford CF, Fidler C *et al*: **Gene expression profiles of CD34+ cells in myelodysplastic syndromes: involvement of interferon-stimulated genes and correlation to FAB subtype and karyotype**. *Blood* 2006, **108**(1):337-345.
8. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M *et al*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome**. *Nature* 2008, **456**(7218):66-72.
9. Bejar R, Stevenson K, Abdel-Wahab O, Galili N, Nilsson Br, Garcia-Manero G, Kantarjian H, Raza A, Levine RL, Neuberg D *et al*: **Clinical Effect of Point Mutations in Myelodysplastic Syndromes**. *New England Journal of Medicine* 2011, **364**(26):2496-2506.
10. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, Pellagatti A, Wainscoat JS, Hellstrom-Lindberg E, Gambacorti-Passerini C *et al*: **Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts**. *New England Journal of Medicine* 2011, **365**(15):1384-1395.
11. Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M *et al*: **Frequent pathway mutations of splicing machinery in myelodysplasia**. *Nature* 2011, **478**(7367):64-69.
12. Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE *et al*: **Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes**. *Nat Genet* 2012, **44**(1):53-57.
13. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB *et al*: **Somatic mutations affect key pathways in lung adenocarcinoma**. *Nature* 2008, **455**(7216):1069-1075.
14. Cancer Genome Atlas Research N: **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061-1068.
15. Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM: **Identifying dysregulated pathways in cancers from pathway interaction networks**. *BMC bioinformatics* 2012, **13**:126.

16. Liu X, Liu ZP, Zhao XM, Chen L: **Identifying disease genes and module biomarkers by differential interactions**. *Journal of the American Medical Informatics Association : JAMIA* 2012, **19**(2):241-248.

17. Zhao XM, Wang RS, Chen L, Aihara K: **Uncovering signal transduction networks from high-throughput data by integer linear programming**. *Nucleic acids research* 2008, **36**(9):e48.

18. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. *Nat Med* 2004, **10**(8):789-799.

19. Zhao J, Zhang S, Wu LY, Zhang XS: **Efficient methods for identifying mutated driver pathways in cancer**. *Bioinformatics* 2012, **28**(22):2940-2947.

20. Vandin F, Upfal E, Raphael BJ: **De novo discovery of mutated driver pathways in cancer**. *Genome Res* 2012, **22**(2):375-385.

21. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: **Predicting the Functional Effect of Amino Acid Substitutions and Indels**. *PloS one* 2012, **7**(10).

22. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protocols* 2009, **4**(8):1073-1081.

23. Brkanac Z, Spencer D, Shendure J, Robertson PD, Matsushita M, Vu T, Bird TD, Olson MV, Raskind WH: **IFRD1 is a candidate gene for SMNA on chromosome 7q22-q23**. *American journal of human genetics* 2009, **84**(5):692-697.

24. Gu Y, Harley IT, Henderson LB, Aronow BJ, Vietor I, Huber LA, Harley JB, Kilpatrick JR, Langefeld CD, Williams AH *et al*: **Identification of IFRD1 as a modifier gene for cystic fibrosis lung disease**. *Nature* 2009, **458**(7241):1039-1042.

25. Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, Stinson J, Forrest WF, Bazan JF, Seshagiri S *et al*: **Distinguishing Cancer-Associated Missense Mutations from Common Polymorphisms**. *Cancer Research* 2007, **67**(2):465-473.

26. Riento K, Ridley AJ: **Rocks: multifunctional kinases in cell behaviour**. *Nat Rev Mol Cell Biol* 2003, **4**(6):446-456.

27. Bottero V, Withoff S, Verma IM: **NF-kappaB and the regulation of hematopoiesis**. *Cell Death Differ* 2006, **13**(5):785-797.

28. Micheli L, Leonardi L, Conti F, Maresca G, Colazingari S, Mattei E, Lira SA, Farioli-Vecchioli S, Caruso M, Tirone F: **PC4/Tis7/IFRD1 stimulates skeletal muscle regeneration and is involved in myoblast differentiation as a regulator of MyoD and NF-kappaB**. *J Biol Chem* 2011, **286**(7):5691-5707.

29. Williams AB, Nguyen B, Li L, Brown P, Levis M, Leahy D, Small D: **Mutations of FLT3/ITD confer resistance to multiple tyrosine kinase inhibitors**. *Leukemia* 2013, **27**(1):48-55.

30. Yamamoto Y, Kiyoi H, Nakano Y, Suzuki R, Kodera Y, Miyawaki S, Asou N, Kuriyama K, Yagasaki F, Shimazaki C *et al*: **Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies**. *Blood* 2001, **97**(8):2434-2439.

31. Takahashi S: **Downstream molecular pathways of FLT3 in the pathogenesis of acute myeloid leukemia: biology and therapeutic implications**. *J Hematol Oncol* 2011, **4**:13.

32. Mjahed H, Girodon F, Fontenay M, Garrido C: **Heat shock proteins in hematopoietic malignancies**. *Exp Cell Res* 2012, **318**(15):1946-1958.

33. White CD, Khurana H, Gnatenko DV, Li Z, Odze RD, Sacks DB, Schmidt VA: **IQGAP1 and IQGAP2 are reciprocally altered in hepatocellular carcinoma**. *BMC Gastroenterol* 2010, **10**:125.

34. Futterer A, Campanero MR, Leonardo E, Criado LM, Flores JM, Hernandez JM, San Miguel JF, Martinez AC: **Dido gene expression alterations are implicated in the induction of hematological myeloid neoplasms**. *J Clin Invest* 2005, **115**(9):2351-2362.

35. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC)**. In: *Current Protocols in Human Genetics.* John Wiley & Sons, Inc.; 2001.

36. Shah SP, Kobel M, Senz J, Morin RD, Clarke BA, Wiegand KC, Leung G, Zayed A, Mehl E, Kalloger SE *et al*: **Mutation of FOXL2 in granulosa-cell tumors of the ovary**. *N Engl J Med* 2009, **360**(26):2719-2729.

37. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

39. Deng X: **SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data**. *BMC Bioinformatics* 2011, **12**(1):267.

40. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm**. *Nat Protoc* 2009, **4**(7):1073-1082.

41. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP: **Predicting the functional effect of amino acid substitutions and indels**. *PloS one* 2012, **7**(10):e46688.

42. Goldberg DE: **Genetic Algorithms in Search, Optimization and Machine Learning**: Addison-Wesley Longman Publishing Co., Inc. ; 1989.

43. Zhao J, Zhang S, Wu L-Y, Zhang X-S: **Efficient methods for identifying mutated driver pathways in cancer**. *Bioinformatics* 2012.

Figures

Figure 1: Flowchart of NGS-based mutation analysis. The pipeline consists of cell preparation, raw data generation, calling mutations, and functional analysis. Each part is described in the Materials and Methods section.

Figure 2: Chromatograms of three validated mutations in IFRD1, IQGAP2 and DIDO1 (a, c, e: forward; b, d, f: reverse). Since these are all heterozygous mutations, at each mutation point there are two peaks of different colors. IFRD1 is in plus strand and the mutation is T to C; a mixed signal of C (blue) and T (red) is seen in a) and of G (black) and A (green) in b). IQGAP2 is in plus strand and the mutation is G to T; a mixed signal of T (red) and G (black) is seen in c) and a mixed signal of A (green) and C (blue) in d). DIDO1 is in minus strand and the mutation is G to A; a mixed signal of A (green) and G (black) is seen in f) and a mixed signal of T (red) and C (blue) in e).

Figure 3: Details of mutations in ras GTPase-activation-like protein (IQGAP2) and Rho-associated coiled-coil kinases 2 (ROCK2). IQGAP2 contains four domains. The mutation (Q1146H) was mapped onto the GTPase activation domain. ROCK2 also contains four domains. The mutation (Y285H) was mapped onto the kinase domain. RBD: Rho-binding domain; PH: Pleckstrin homology domain; CRD: cysteine-rich domain.

Figure 4: Two networks predicted by IPA. A. The first was identified with cancer, hematologic disease, and molecular transport. 26 genes that we input were annotated as gray and red and 11 red nodes presented genes came from our mutation detection pipeline. This network centered at CDKN2A and NFκB, and the mutated genes are very close to them; B. The second network was identified with cancer, hematologic disease, and cellular development. This network contained 19 genes that we input (grey and red). Three (red) were found by the mutation detection pipeline. This network centered at ERK1/2.

Figure 5: Two mutated driver pathways. A. The results of our RMDP$^{\text{scoring}}$ method when k was set as 6. Two optimal gene sets were output. Green boxes indicate mutations. Each matrix involves six patients who had exclusive mutations which are all in the pathway in cancer. This driver pathway is associated with cancer. B. The figure shows results of our RMDPscoring method when k was set as 4. Two optimal gene sets were output. Green boxes indicate mutations. Each matrix involves five patients and they have exclusive mutations which are all involved in the regulation of actin cytoskeleton. This mutated driver pathway is associated with regulation of the actin cytoskeleton.

Figure 6: Functional profiles of genes enriched in the cancer driver pathway. SMAD4 and MLH1 are regulated by FLT3 through the MAPK9 pathway, affect transcription, and lead to anti-apoptosis, abnormal cell growth, and differentiation blockage. Hsp90 chaperones FLT3 and the downstream signaling molecules and is also involved in this regulation and following abnormal hematopoiesis.

Tables

Table 1: Mutations detected in 20 MDS samples, with the threshold of coverage set as 30. POS: the position of mutated base in chromosome; ref: information used as a reference; mut: mutation in our sample; het: heterozygous mutation; hom: homozygous mutation; PROVEN: prediction of PROVEAN; SIFT: prediction of SIFT; NA: not available; * in the Amino acid column refers to a nonsense nutation.

Tables

Table 1: Mutations detected in 20 MDS samples, with the threshold of coverage set as 30

POS: the position of mutated base in chromosome; ref: information used as a reference; mut: mutation in our sample; het: heterozygous mutation; hom: homozygous mutation; PROVEN: prediction of PROVEAN; SIFT: prediction of SIFT; NA: not available; * in the Amino acid column refers to a nonsense nutation.

Symbol	POS	Base	Amino acid	Type	PROVEAN	SIFT
		(ref/mut)	(ref/mut)			
AHNAK	62292425	A/G	M/T	het	Deleterious	Tolerated
AHNAK	62298107	T/G	K/T	het	Deleterious	Damaging
BDP1	70763215	T/C	C/R	het	Deleterious	Tolerated
C10orf12	98743732	T/A	M/K	het	Deleterious	Damaging
C11orf84	63585765	C/T	L/F	hom	Neutral	Damaging
CALR	13054721	G/T	E/D	het	Neutral	Damaging
CEP135	56878014	C/T	R/C	het	Deleterious	Damaging
CMTM7	32483465	G/A	R/H	hom	Deleterious	Damaging
COX7C	85913936	G/A	G/S	het	Deleterious	Tolerated
DEK	18258280	T/G	K/N	het	Neutral	Damaging
DEK	18258281	T/G	K/R	het	Neutral	Damaging
DEK	18258282	T/G	K/Q	het	Neutral	Damaging
DIDO1	61538644	G/A	S/L	het	Deleterious	Damaging
EIF3A	120801799	C/A	G/V	het	Neutral	Damaging
ERAP2	96228086	G/T	D/Y	het	Deleterious	Damaging
GON4L	155746190	A/C	F/C	hom	Deleterious	Tolerated
GON4L	155746202	G/C	T/S	hom	Deleterious	Tolerated
GON4L	155746206	T/G	T/P	het	Neutral	Damaging
HSD17B4	118872214	A/G	E/G	het	Deleterious	Damaging
IDH1	209113113	G/T	R/S	het	Deleterious	Damaging
IFFO2	19246218	T/C	E/G	het	Deleterious	Damaging
IFRD1	112097081	T/C	C/R	het	Deleterious	Damaging
IQGAP2	75970445	G/T	Q/H	het	Deleterious	Damaging
MSL2	135870569	A/T	I/K	het	Neutral	Damaging
MUC12	100645785	C/T	P/S	het	Neutral	Damaging
NRD1	52301829	G/A	P/L	het	Deleterious	Tolerated
NUCB2	17336970	A/T	E/V	het	Deleterious	Damaging
NUMB	73822386	C/T	W/*	het	Nonsense	Nonsense
NXF1	62567942	T/C	E/G	het	Deleterious	Damaging
PARP2	20822340	T/A	C/S	het	Deleterious	Damaging
PHF20L1	133851794	T/C	L/P	het	Deleterious	Damaging
PLAC8	84015929	C/A	D/Y	het	Deleterious	Damaging
PPHLN1	42745802	A/G	E/G	het	Deleterious	Damaging

RAD21	117875473	C/T	G/E	het	Deleterious	Damaging
RBBP6	24578657	G/A	A/T	het	Neutral	Damaging
ROCK2	11367395	A/G	Y/H	het	Deleterious	Damaging
RPS3	75113469	T/C	L/P	het	Deleterious	Damaging
SCP2	53516318	T/C	M/T	het	Deleterious	Damaging
SP100	231307768	T/G	L/R	het	Deleterious	Damaging
SP100	231307771	A/C	E/A	het	Deleterious	Damaging
TSPYL2	53115453	A/G	I/V	het	Neutral	Damaging
UBE2L3	21965309	A/G	K/R	het	Deleterious	Damaging
YLPM1	75295994	T/C	L/P	het	Deleterious	NA
ZCCHC11	52954588	C/T	W/*	het	Nonsense	Nonsense

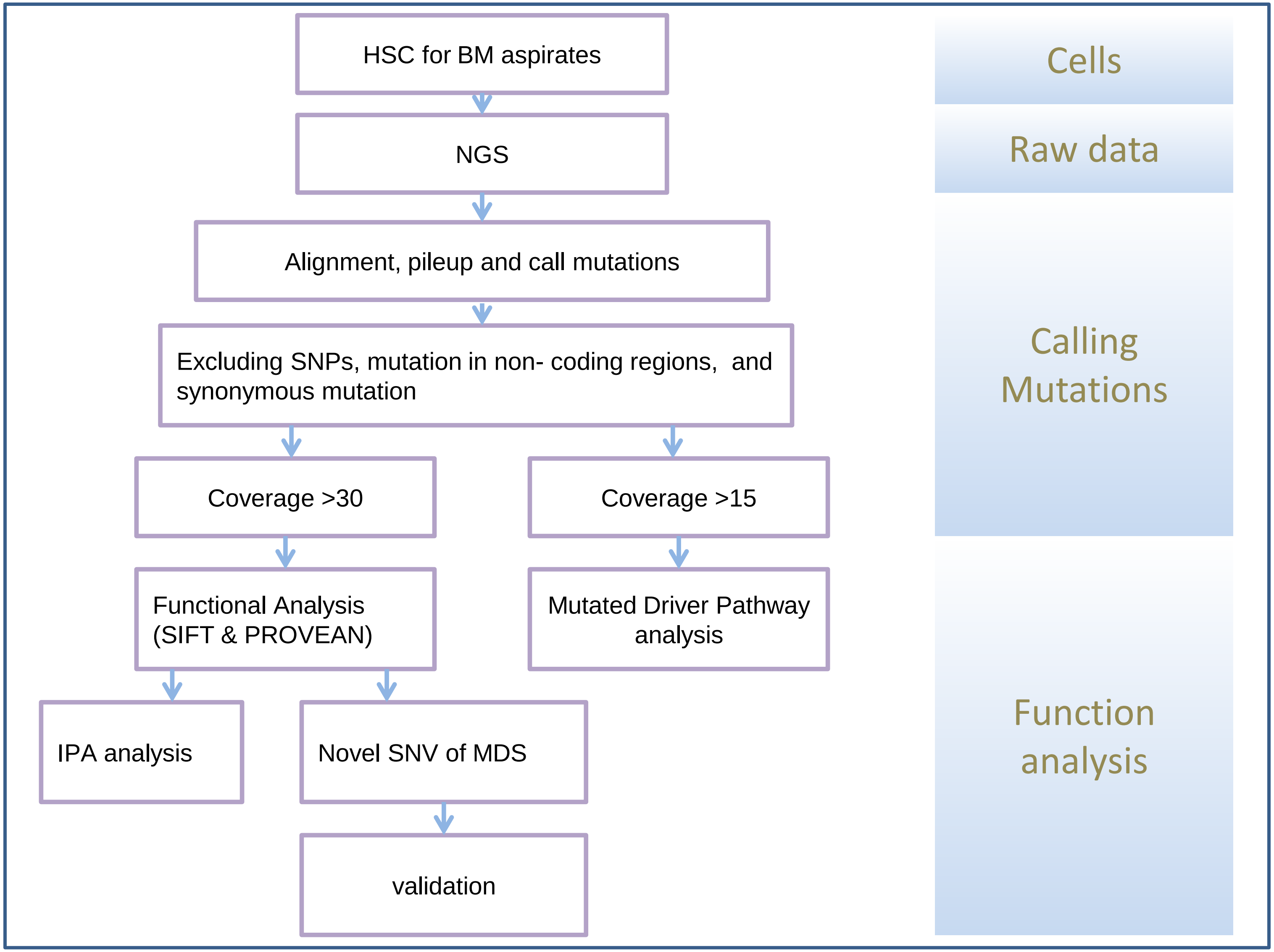

HSC for BM aspirates
NGS
Alignment, pileup and call mutations
Excluding SNPs, mutation in non- coding regions, and synonymous mutation
Coverage >30
Coverage >15
Functional Analysis (SIFT & PROVEAN)
Mutated Driver Pathway analysis
IPA analysis
Novel SNV of MDS
validation
Cells
Raw data
Calling Mutations
Function analysis

Figure 1

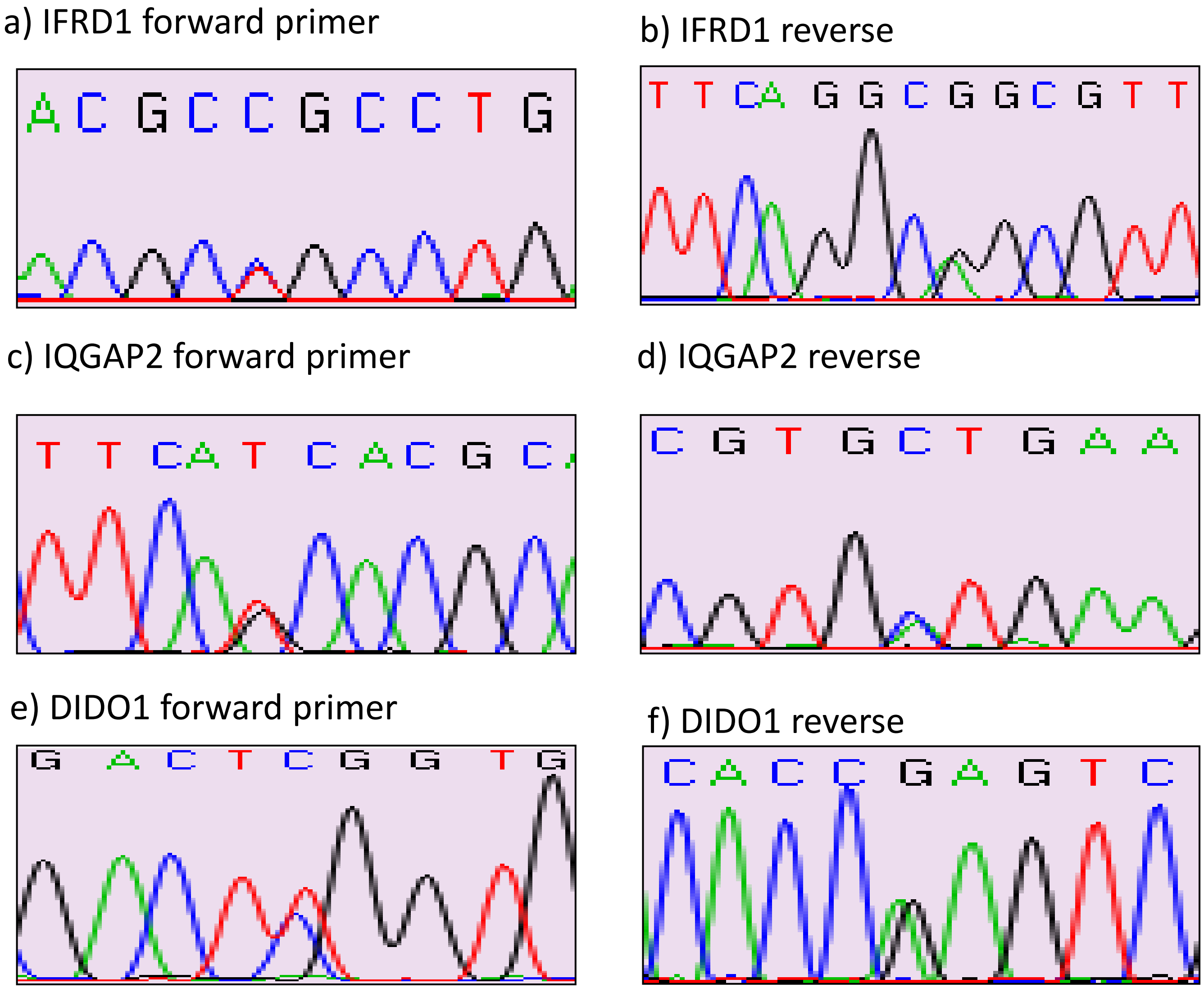
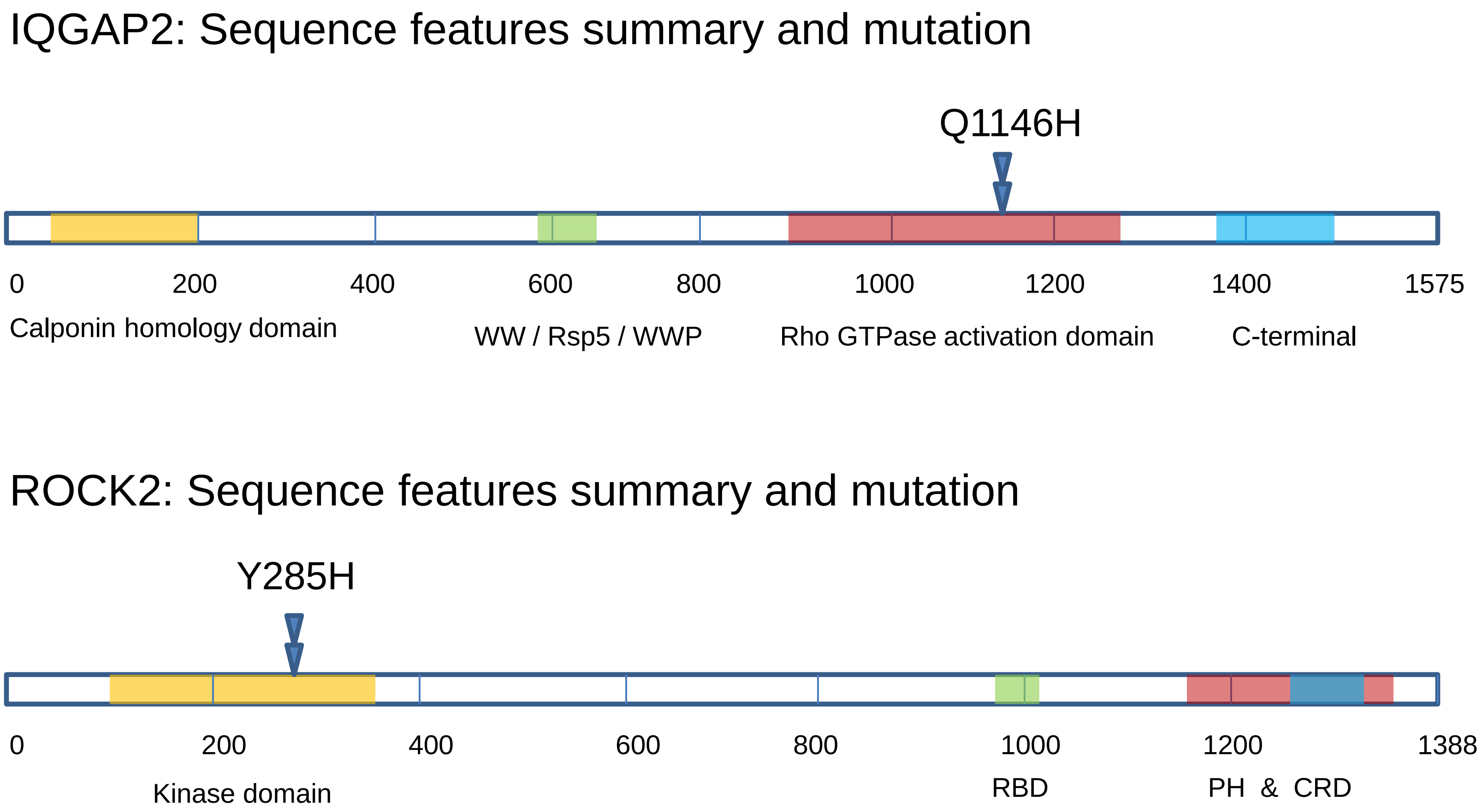

a) IFRD1 forward primer
A C G C C G C C T G
b) IFRD1 reverse
T T C A G G C G G C G T T
c) IQGAP2 forward primer
T T C A T C A C G C
d) IQGAP2 reverse
C G T G C T G A A
e) DIDO1 forward primer
G A C T C G G T G
f) DIDO1 reverse
C A C C G A G T C

Figure 2



IQGAP2: Sequence features summary and mutation

ROCK2: Sequence features summary and mutation

Figure 3

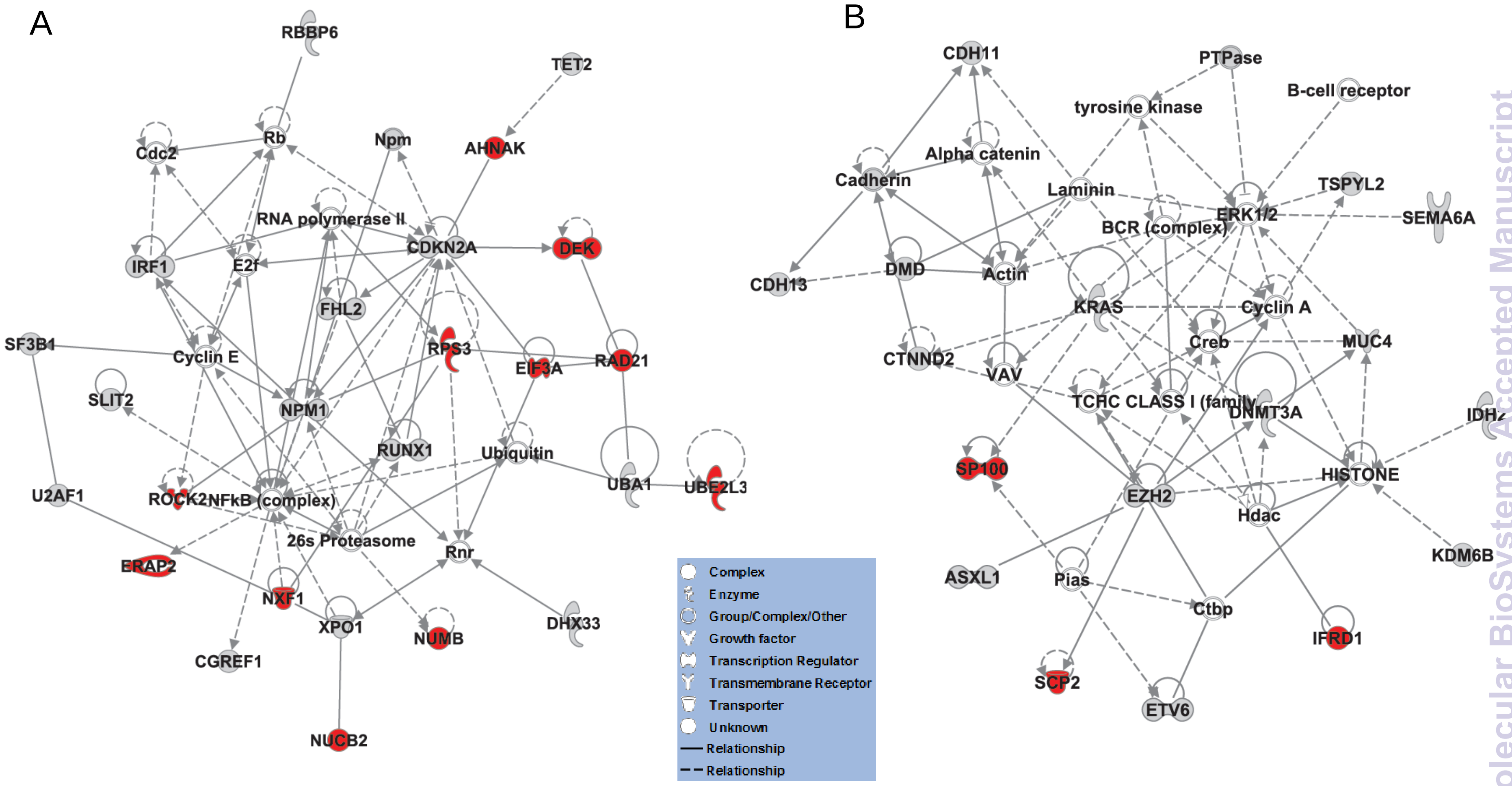

A
RBBP6
TET2
Rb
Npm
AHNAK
Cdc2
RNA polymerase II
CDKN2A
DEK
IRF1
E2f
FHL2
SF3B1
Cyclin E
RPS3
EIF3A
RAD21
SLIT2
NPM1
RUNX1
Ubiquitin
UBA1
UBE2L3
U2AF1
ROCK2
NFkB (complex)
26s Proteasome
Rnr
ERAP2
NXF1
XPO1
NUMB
DHX33
CGREF1
NUCB2
B
CDH11
PTPase
B-cell receptor
tyrosine kinase
Alpha catenin
Cadherin
Laminin
TSPYL2
ERK1/2
SEMA6A
BCR (complex)
CDH13
DMD
Actin
KRAS
Cyclin A
Creb
MUC4
CTNND2
VAV
TCHC CLASS I (family)
DNMT3A
IDH2
SP100
EZH2
HISTONE
Hdac
KDM6B
ASXL1
Pias
Ctbp
IFRD1
SCP2
ETV6
Complex
Enzyme
Group/Complex/Other
Growth factor
Transcription Regulator
Transmembrane Receptor
Transporter
Unknown
Relationship
Relationship

Figure 4

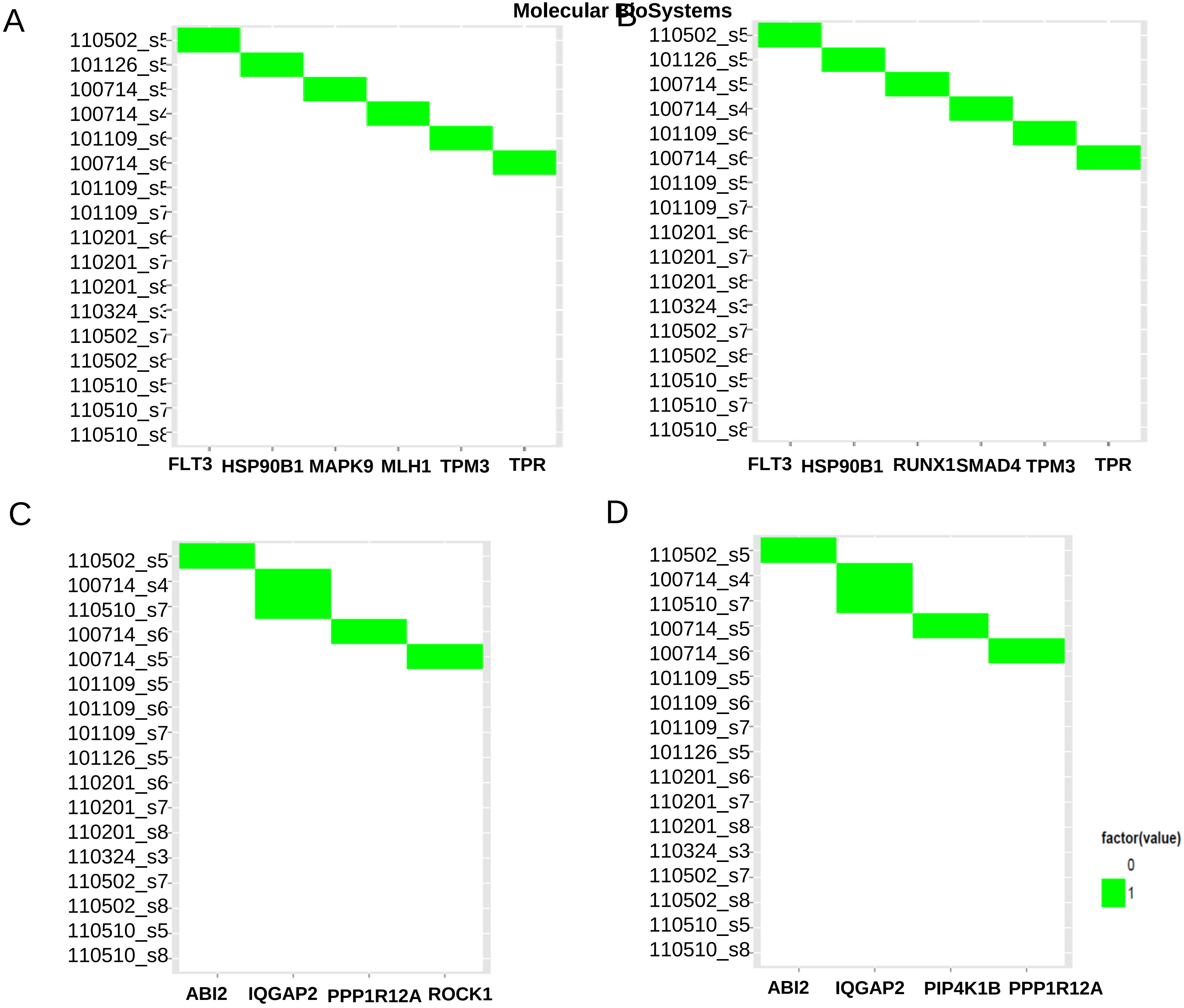

A
110502_s5
101126_s5
100714_s5
100714_s4
101109_s6
100714_s6
101109_s5
101109_s7
110201_s6
110201_s7
110201_s8
110324_s3
110502_s7
110502_s8
110510_s5
110510_s7
110510_s8
FLT3
HSP90B1
MAPK9
MLH1
TPM3
TPR
B
110502_s5
101126_s5
100714_s5
100714_s4
101109_s6
100714_s6
101109_s5
101109_s7
110201_s6
110201_s7
110201_s8
110324_s3
110502_s7
110502_s8
110510_s5
110510_s7
110510_s8
FLT3
HSP90B1
RUNX1
SMAD4
TPM3
TPR
C
110502_s5
100714_s4
110510_s7
100714_s6
100714_s5
101109_s5
101109_s6
101109_s7
101126_s5
110201_s6
110201_s7
110201_s8
110324_s3
110502_s7
110502_s8
110510_s5
110510_s8
ABI2
IQGAP2
PPP1R12A
ROCK1
D
110502_s5
100714_s4
110510_s7
100714_s5
100714_s6
101109_s5
101109_s6
101109_s7
101126_s5
110201_s6
110201_s7
110201_s8
110324_s3
110502_s7
110502_s8
110510_s5
110510_s8
ABI2
IQGAP2
PIP4K1B
PPP1R12A
factor(value)
0
1

Figure 5

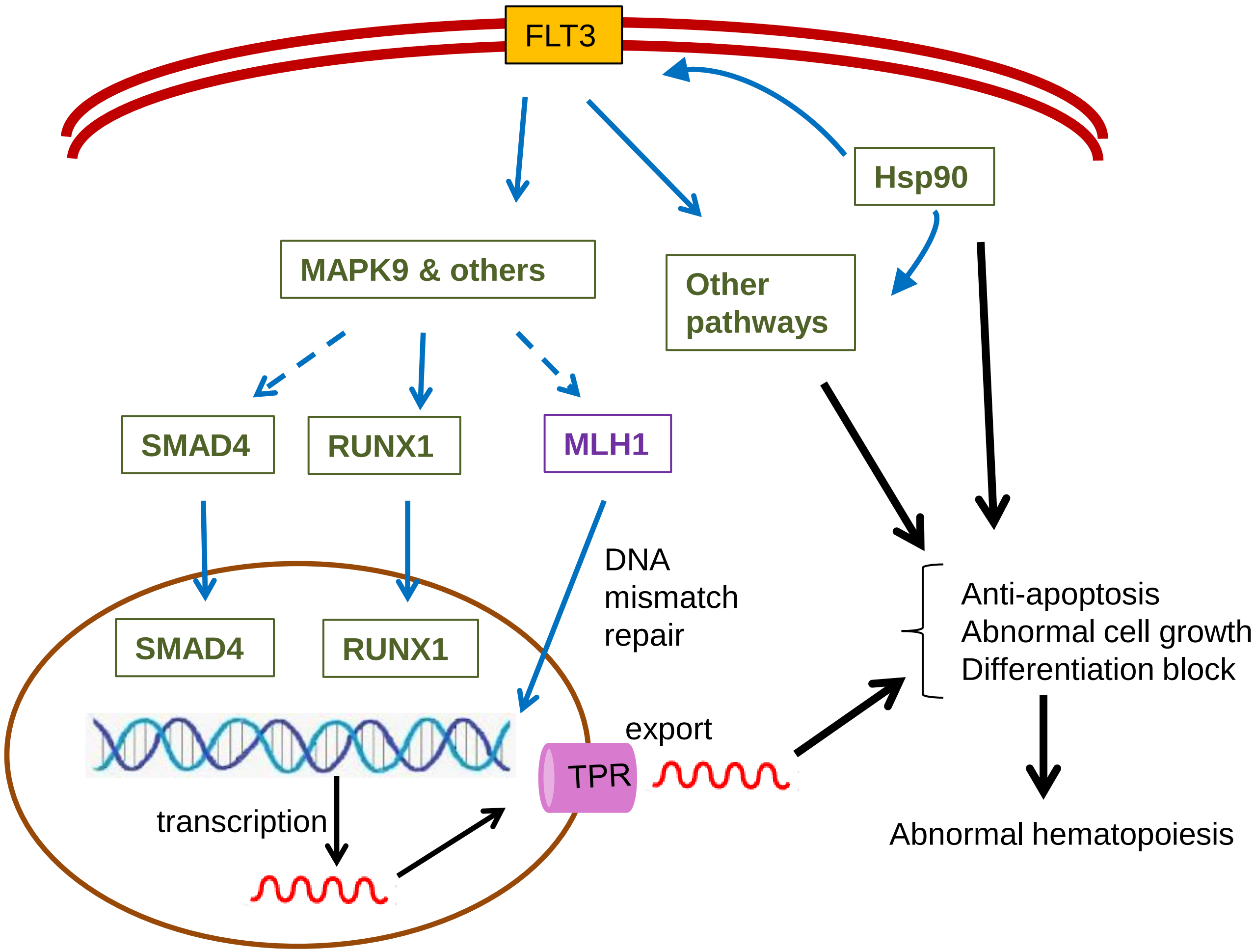

FLT3
Hsp90
MAPK9 & others
Other pathways
SMAD4
RUNX1
MLH1
DNA mismatch repair
SMAD4
RUNX1
Anti-apoptosis
Abnormal cell growth
Differentiation block
export
TPR
transcription
Abnormal hematopoiesis

Figure 6