UDC 57.087.1::616-006.4
doi:10.31799/1684-8853-2020-6-50-59

Articles

Building and evaluation of bioinformatic pipeline for determination of clonal profiles in myelodysplastic syndrome

D. S. Bug[a], *M. D., orcid.org/0000-0002-5849-1311, dmitriybs@1spbgmu.ru*
A. A. Prikhodko[a], *Student, orcid.org/0000-0002-0001-7932*
E. A. Bakin[a,b], *PhD, Tech., Associate Professor, orcid.org/0000-0002-5694-4348*
A. V. Tishkov[a], *PhD, Phys.-Math., Associate Professor, orcid.org/0000-0002-4282-8717*
N. V. Petukhova[a], *PhD, Biol., orcid.org/0000-0001-6397-824X*
I. M. Barkhatov[a], *PhD, Med., orcid.org/0000-0002-8000-3652*
E. V. Morozova[a], *PhD, Med., Associate Professor, orcid.org/0000-0002-9605-485X*
I. S. Moiseev[a], *Dr. Sc., Med., Associate Professor, orcid.org/0000-0002-4332-0114*
[a]*Pavlov First Saint Petersburg State Medical University, 6-8, L'va Tolstogo St., 197022, Saint-Petersburg, Russian Federation*
[b]*Saint-Petersburg State University of Aerospace Instrumentation, 67, B. Morskaia St., 190000, Saint-Petersburg, Russian Federation*

Introduction: *There is growing evidence of a connection between tumor clonal profile and its clinical impact. However, there is a lack of a feasible and reliable method for clonal profiling in actual clinical practice. Myelodysplastic syndrome is a clonal hematopoietic stem cell disorder characterized by morphological dysplasia, cytopenia and a high risk of evolution to acute myeloid leukemia. The clinical outcome of myelodysplastic syndrome is greatly heterogeneous; therefore, specific examination of clonal profiles is needed to resolve the prognosis of patients with such complex disorders.* ***Purpose:*** *Development of a pipeline specifically for determining the clonal profiles in patients with myelodysplastic syndrome on the basis of target next-generation sequencing data.* ***Results:*** *The pipeline was developed and evaluated on a set of 35 patients with high-risk myelodysplastic syndrome. It is possible to use the target sequencing data in order to assess the heterogeneity of clonal profiles and characterize their genetic features. This approach allows you to identify the consistency between a specific individual profile and the disease prognosis, which can be critical for the treatment decision. Herein, the characterization and analysis of clonal profiles are presented.* ***Practical relevance:*** *The information about relation patterns between clonal profile characteristics (number of subclones, mutations-per-clone rate) and clinical outcome can be used by doctors in current practice for a more accurate therapy selection depending on the identified individual specificity of the disease.*

Keywords *— myelodysplastic syndrome, bioinformatics pipeline, clonal profile, primary mutation, subclone, target sequencing, next-generation sequencing.*

For citation: Bug D. S., Prikhodko A. A., Bakin E. A., Tishkov A. V., Petukhova N. V., Barkhatov I. M., Morozova E. V., Moiseev I. S. Building and evaluation of bioinformatic pipeline for determination of clonal profiles in myelodysplastic syndrome. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2020, no. 6, pp. 50–59. doi:10.31799/1684-8853-2020-6-50-59

Introduction

Adult myelodysplastic syndrome (MDS) occurs as a result of the gradual accumulation of somatic mutations [1]. A set of cells derived from a mutated cell is called a clone. Subclones arise from a primary clone (so-called subclones) in the process of acquiring new somatic mutations (tumor progression, clonal evolution), therefore, they are also considered as clones. To predict the tumor development pathway, it is essential to characterize the clone with its subclones [2].

Each clone is defined by its unique mutational profile providing to trace the relation between clone genotype and prognosis of the disease. The presence of subclones with driver mutations in oncogenes has been shown to negatively affect the outcome of the disease in case of chronic lymphocytic leukemia [3]. A correlation between a large number of subclones and an unfavorable prognosis was identified in lung, breast, prostate, kidney tumors, as well as low-grade glioma [4–6].

One of the main approaches of next-generation sequencing (NGS) is targeted sequencing (TS) method focused on specific genes panel in order to determine the current mutation load. High precision, relatively low cost, multiple genes analysis at once are the main advantages of TS. However, there are some technical issues associated with the implementation of this technique in actual clinical practice: lack of verification methods and standard pipelines for analysis of sequencing data, a need to adapt for particular diseases. TS method used with additional bioinformatic tools could be helpful in clonal profile deduction.

Thus, the purpose of present work is to build an analytical pipeline using modern bioinformatic tools specifically adapted for determination of the

clonal profiles in patients with MDS on the basis of TS. Such evaluation might help to derive the relation and corresponding patterns between the number and the genotype of subclones, and subsequent disease prognosis.

Materials and methods

1. Group selection.

Next-generation sequencing of 35 patients with high-risk MDS was performed (Table 1) [7].

2. DNA sequencing.

Genomic DNA was extracted from bone marrow using TriZ reagent extraction Kit (Inogene, Russia) and stored at −80 °C. The quality of the samples was analyzed with Qubit 4.0 (Thermo Fisher, CA, USA). The libraries for target sequencing of the genes were prepared using KAPA HyperPlus Kit (Roche, Switzerland). The enrichment of targeted genome sequences was performed using SeqCap EZ Target Enrichment System (Roche, Switzerland). Sequencing was performed by MiSeq benchtop sequencer using MiSeq Reagent Kits v2 (Illumina, USA).

■ ***Table 1***. Characteristics of patients before and after filtering of germinal and false positive variants

Parameter	Value	
	before filtering	after filtering
Age median (interval), years	49 (18–80)	46,4 (18–80)
Male/female	21/14	16/11
After bone marrow transplant	25 (71,4%)	20 (74,1%)
Secondary MDS (after prior chemo- or radiotherapy)	5 (14,3%)	5 (18,5%)
Diagnosis		
5q deletion	1 (2,9%)	1 (3,7%)
Excess of blasts I	13 (37,1%)	10 (37,0%)
Excess of blasts II	19 (54,3%)	14 (51,9%)
Multilineage dysplasia	2 (5,7%)	2 (7,4%)
Risk according to IPSS-R		
Very low	0	0
Low	1 (2,9%)	1 (37%)
Intermediate	5 (14,3%)	4 (14,8%)
High	15 (42,8%)	12 (33,3%)
Very high	14 (40,0%)	13 48,1%)

3. Data preprocessing.

The quality of sequencing reads was analyzed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/); adapters were eliminated by Trimomatic 0.39 [8].

4. Alignment, analysis and generation of the detected variants list.

Bioinformatic analysis was performed according to the GATK-4 manual [9], with alignment to the GRCh38 version of the human genome using BWA 0.7.17 [10], and determination of somatic variants was accomplished with Mutect2 4.1.5.0 [11]. The resulting list of variants was annotated using Ensemble Variant Effect Predictor 99 [12] and ANNOVAR [13].

The custom scripts used in the analysis on the basis of GATK Best Practices pipelines, were deposited to the Github portal (https://github.com/bugds/BashGATK). Parallelization was performed using GNU Parallel [14].

5. Filtering of the germinal variants.

When analyzing the identified gene variants, we excluded those variants whose frequency in different populations according to GnomAD [15] exceeds 1%, and the allelic load (taking into account the chimerism) was in the range of 30–70% and 90–100% [16].

6. Removing false positives.

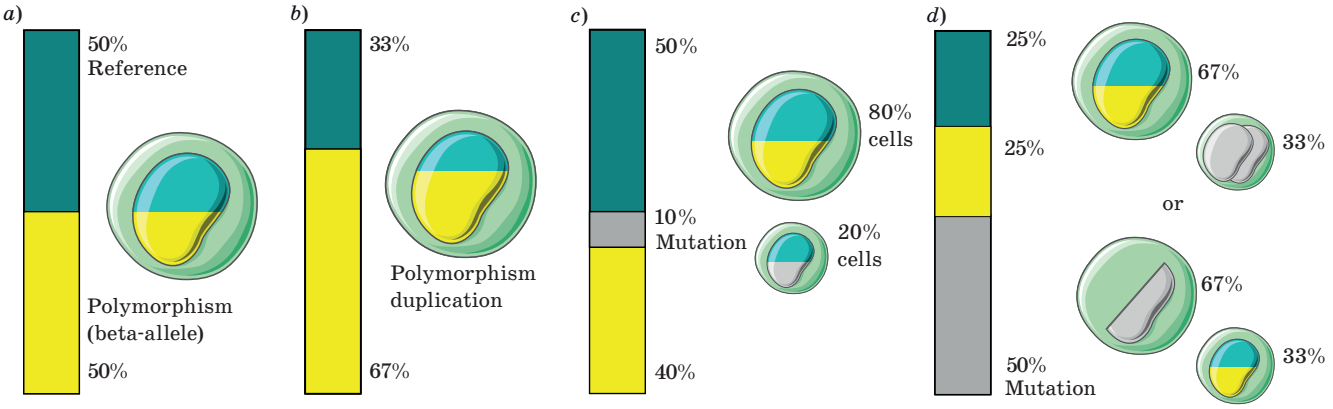
Variants not filtered by Mutect2 were not taken into account. Additionally, variants found in two or more samples with the same allelic load were removed from the analysis that indicate a direct sign of mispriming [17].

Samples were removed from the analysis in cases where less than two variants remained as a result of filtering (the study of tumor progression is based on comparing allelic loads of variants and is possible only if there are two or more variants). Finally, 27 patient samples remained in the study (see Table 1).

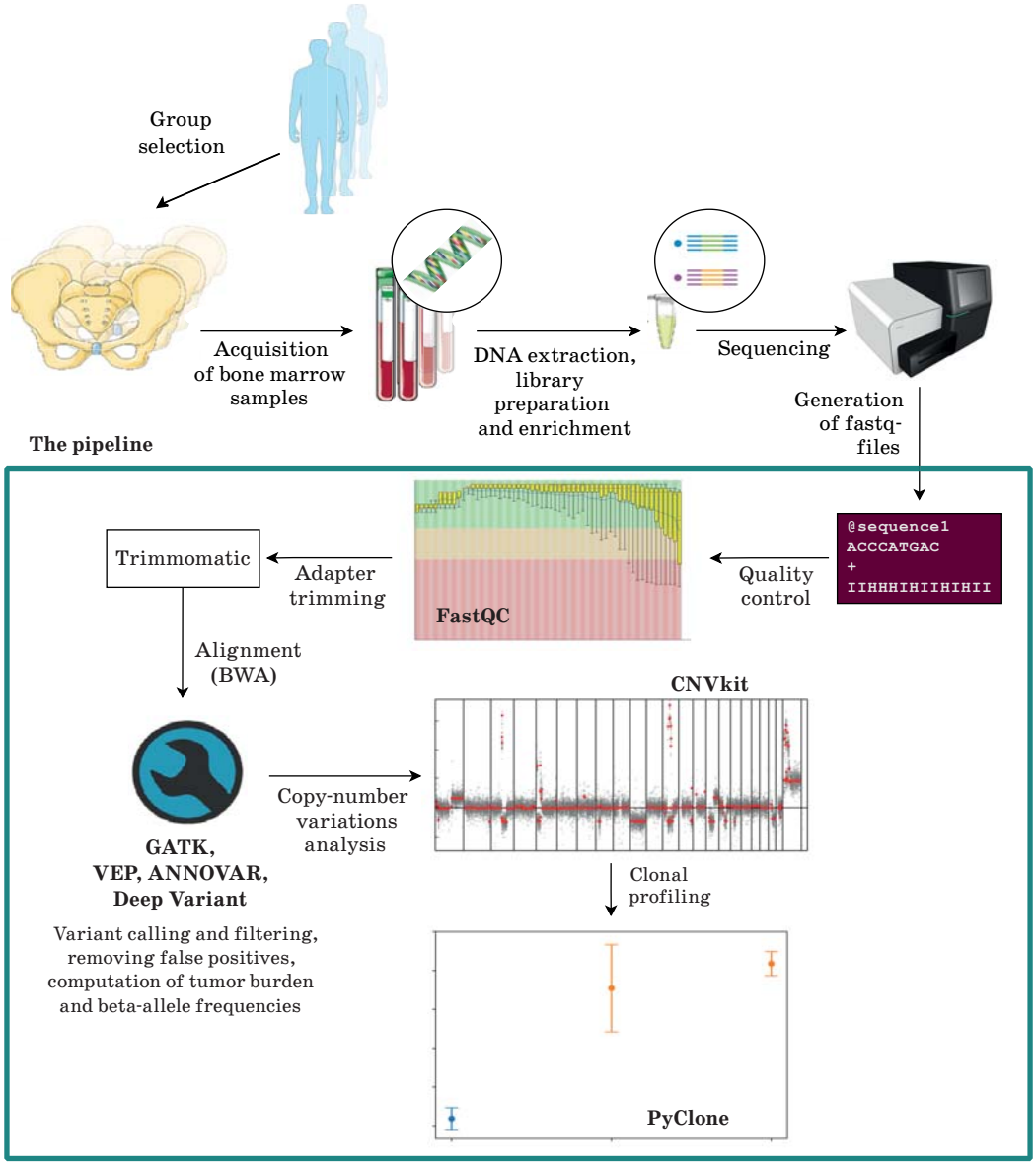
7. Computation of tumor burden and beta-allele frequencies.

Beta-alleles are the polymorphisms detected during sequencing (Fig. 1, *a*). Their allelic load must be taken into account when calculating copy-number variations (see point 8): for instance, a germinal variant in a heterozygous state with an allelic load of about 33 or 67% indicates a duplication of this region (Fig. 1, *b*). The Deep Variant 0.10.0 software tool was used to determine polymorphisms [18].

In each sample, variants with the maximum allelic load were identified — they are primary, and approximately reflect the proportion of tumor cells in the sample. In this case, we assume that the variants with the maximum allelic frequency in each sample are not located in the regions of the genome with a copy-number variation (insertions and deletions), and are also heterozygous. In this case, the pro-



■ ***Fig. 1.*** Compliance between the allelic load detected during sequencing (columns) and the genotype configuration in normal (*a*), with polymorphism duplication (*b*), the presence of a mutation in 20% of cells (*c*) and the 50% allelic load of the mutation (*d*)



■ ***Fig. 2.*** The workflow of clonal profiling

portion of tumor cells can be calculated as the double number of variant allelic loads (Fig. 1, *c*). Cases where variants with maximum loads are located in regions of the genome with copy-number variation require high-attention. For example, if a variant is found in 50% of the reads, this may indicate both a 100% tumor load and the above phenomena associated with a copy-number variation (Fig. 1, *d*). Similar phenomena can be suspected when studying allelic loads of the other variants found in the same sample. They were analyzed manually, and if interpretation was not possible, they were excluded (it was considered that there is copy-number variation when identifying stable change of reading depth in the region with an allele mutation with the maximum frequency, and the mutation with the maximum allelic frequency cannot be considered primary).

8. **Copy-number variations analysis.**

To determine the copy-number variation, CNVkit 0.9.5 was used, an algorithm that utilizes both target gene reads and non-specifically captured non-target reads to identify the copy-numbers evenly across the entire genome [19]. This tool has a relatively high accuracy [20] and is designed specifically for detecting acquired copy-number variation together with TS. To study somatic insertions and deletions, it is necessary to take into account the proportion of tumor cells in the sample determined at the previous stage (see point 7). The copy-number is calculated by comparing the reading depth in certain positions in the control and pathological samples, taking into account the proportion of tumor cells and the size of the beta allele. In CNVkit, a pooled or single reference can be used as a reference material. In this study, the usage of a sample with more than 99% chimerism and a normal karyotype was chosen as a reference, which corresponds to complete cytogenetic remission taken from a patient 4 weeks after bone marrow transplantation.

9. **Clonal profiling.**

PyClone 0.13.1 software tool was used to determine the clonal profiles [21] — the algorithm performs Bayesian clustering method to group somatic mutations into assumed clonal clusters with an assessment of their cellular prevalence (the proportion of affected cells) and taking into account the allelic imbalances introduced by copy-number variations and contamination of normal cells.

The complete workflow of analysis is presented in Fig. 2.

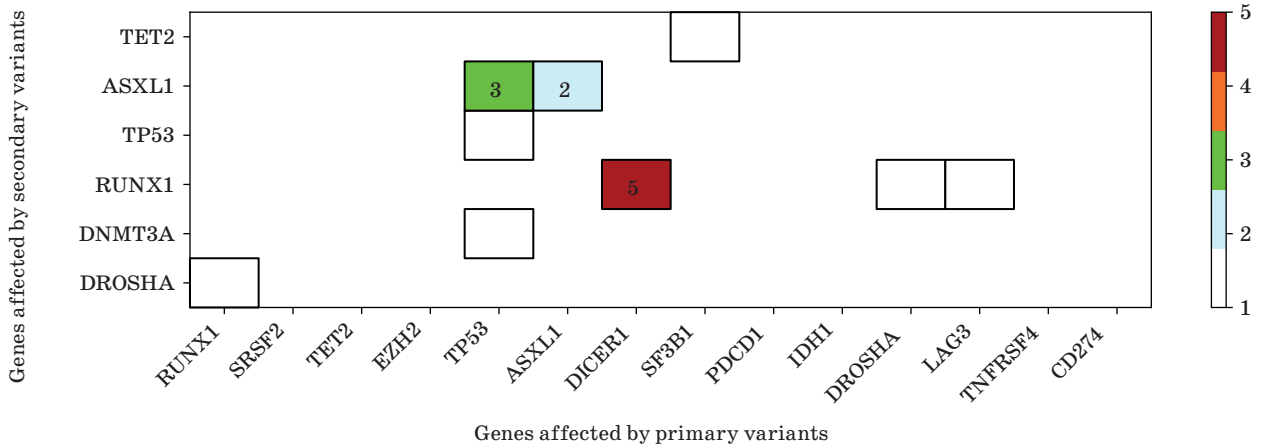
Results

The whole analysis took 7 hours 2 minutes (Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, 32G).

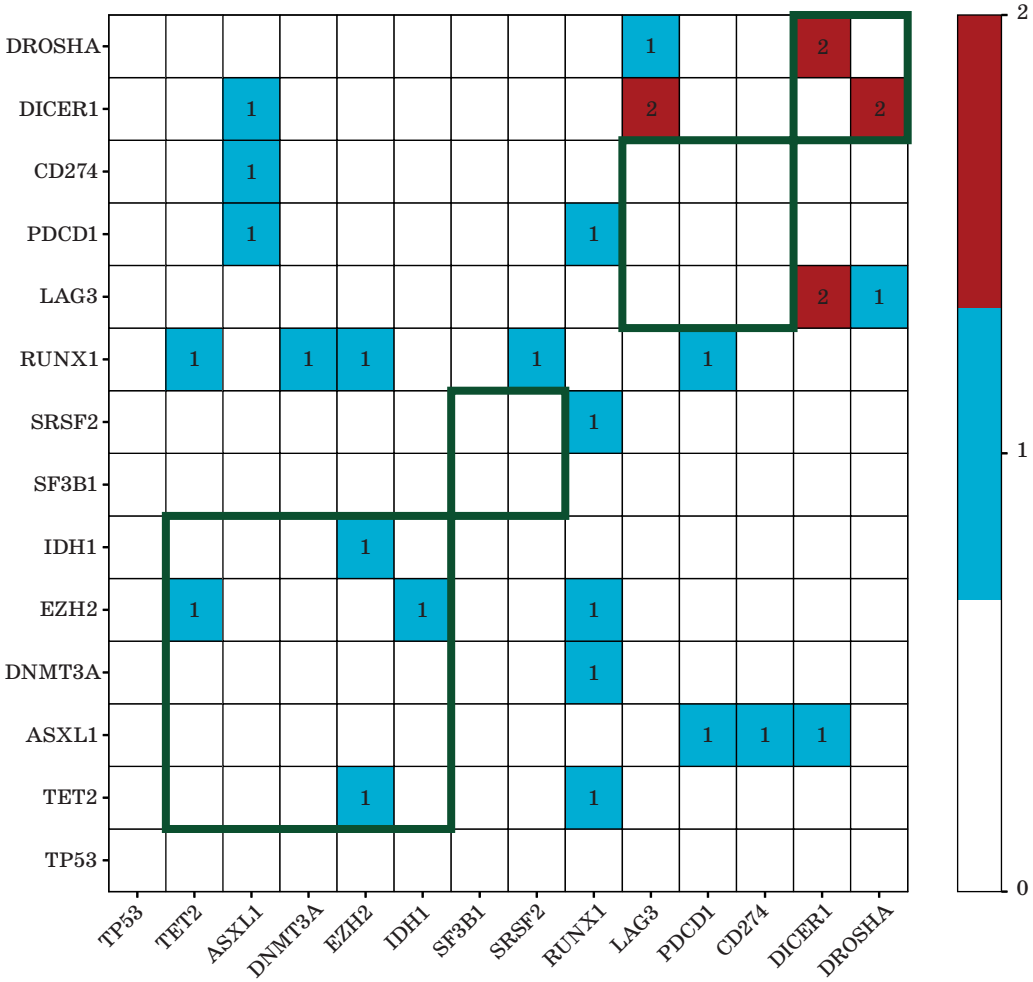
Under the analysis, one of the main issues was the lack of an out-of-the-box tool for reformatting the CNVkit output format into the one suitable for PyClone. However, our pipeline avoids the majority of such complications by using a single package (GATK) for the most part of the analysis.

Mutational profiles of the patients were derived in form of tables; each variant was annotated with its original allele frequency, the calculated tumor load (i. e. a fraction of cells where this variant was observed), the standard deviation of the tumor load, and the cluster-ID that included the variant itself, based on the calculated tumor burden. Two matrices were obtained based on the patients' clonal profiles: one, depicting the co-occurrence of primary and secondary variants from the genes perspective (Fig. 3), and another, describing genes affected by mutations in the same clusters (Fig. 4).

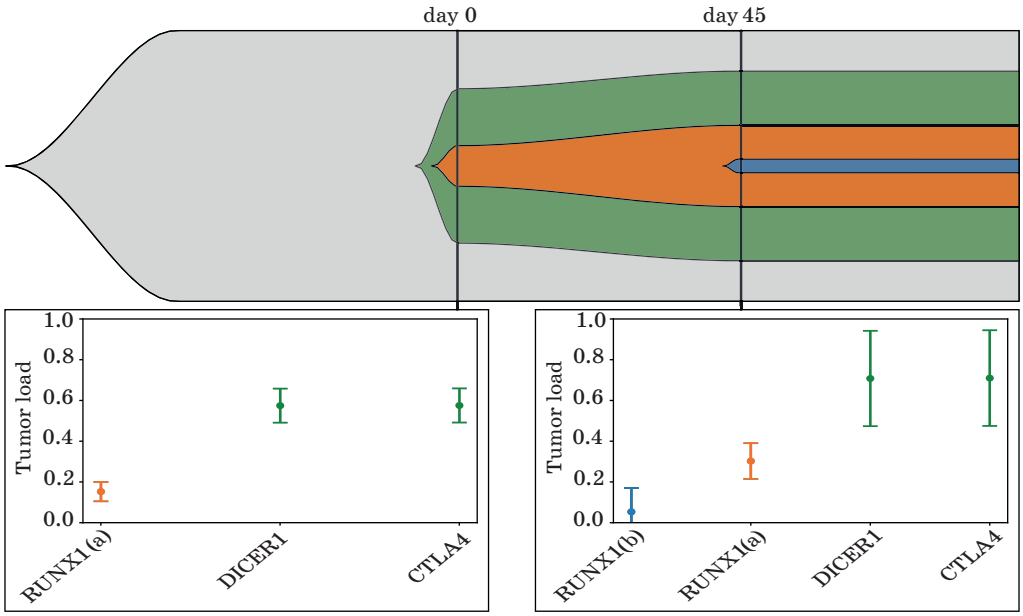
The dynamics of tumor progression was described for a patient, who was sequenced twice with samples taken in 45 days (Fig. 5 [22]). A subclone with *RUNX1* mutation overcame the detection



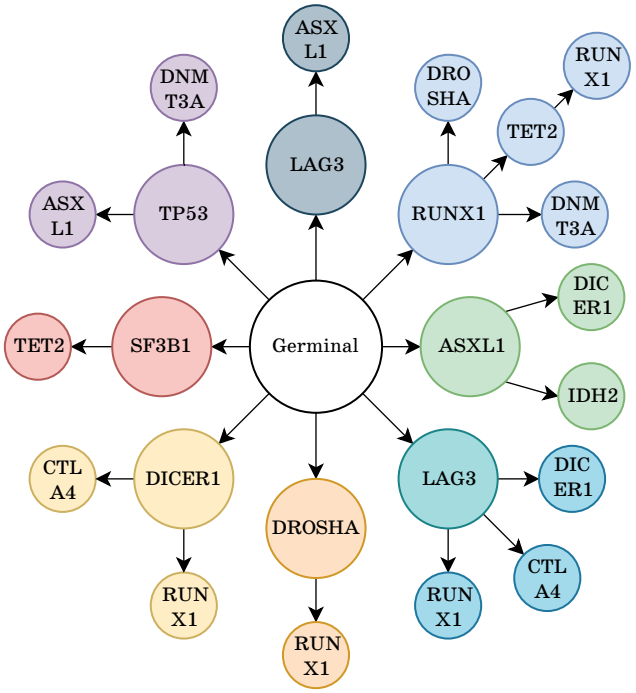
■ ***Fig. 3.*** Matrix of co-occurrence of primary and secondary gene variants. The numbers *1–5* of such events are differentiated by colors



■ ***Fig. 4.*** Matrix of co-occurrence of gene variants in the same clusters. The numbers *0–2* of identified clusters with a certain combination of variants are represented by different colors. Combinations of variants affecting genes of a single functional group are marked by bold frames



■ ***Fig. 5.*** Dynamics of tumor progression



■ ***Fig. 6.*** Schematic representation of all possible paths of tumor progression for variants with >10% load discovered across all analyzed samples

threshold, and two other clones identified previously have grown.

Discovered paths of tumor progression were plotted (Fig. 6). Moving along the arrows depicts acquiring a mutation in a specific gene by a clone. The diagram demonstrates all variants of tumor progression that may be inferred from the observable clonal profiles.

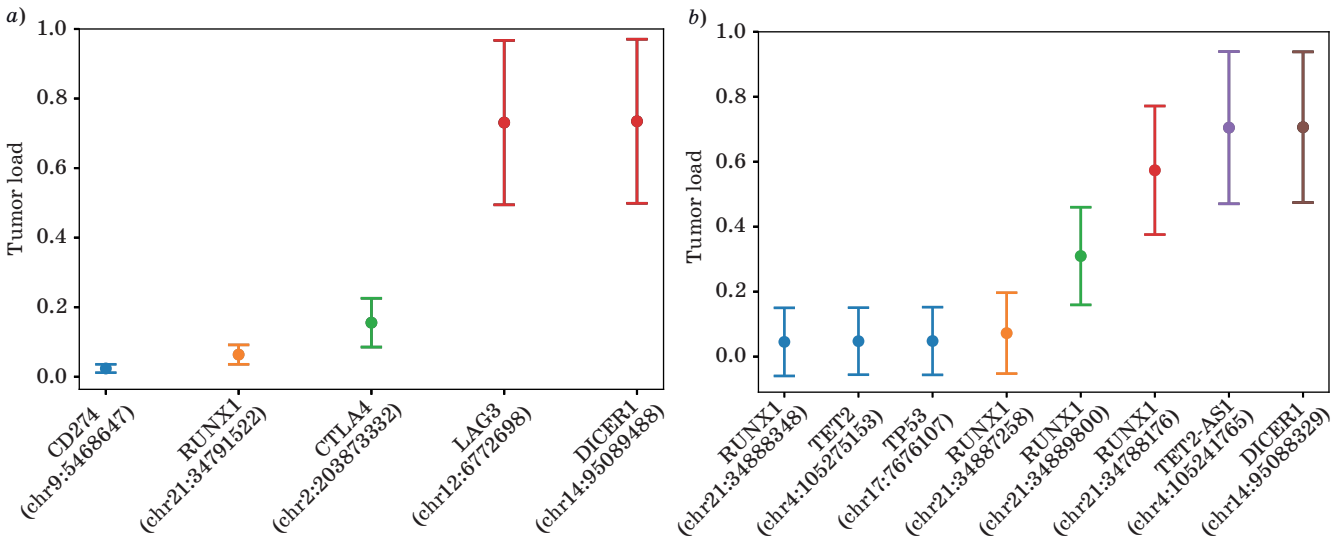
Discussion

We have developed and tested the pipeline for clonal profiles determination in patients with MDS. It is advantageous to identify the correlation between the clonal profile, their genotypic features and disease prognosis, that has been effectively demonstrated for other pathologies [3–6].

Genotypes of clones and subclones were plotted (Fig. 7). Whiskers indicate the standard deviation of tumor burden of a certain variant. The color of a mark shows the affiliation to a specific cluster (clone or subclone). Genes affected by a variant and its genomic position in chromosomes are marked on a horizontal axis.

Five mutations were found in the same sample (Fig. 7, *a*) herewith two variants (in *LAG3*, *DICER1* genes) grouped in a single cluster. This fact, considering their high tumor load (> 50%), can be interpreted as their coexistence in the same cells. Otherwise, such a clonal profile would only be observed if the mutation occurs again in the same position, which is highly improbable.

In the other case (Fig. 7, *b*), eight mutations were discovered in genes: *DICER1*, *TET2-AS1*, *RUNX1*, *TP53*, and *TET2*. Probably, the variant present in the largest number of cells (marked brown, mutation in *DICER1*) occurred almost simultaneously with the mutation in *TET2-AS1*, which led to the neoplasm formation. Then, all the other variants developed within the proliferating tumor. Three mutations located in the same cluster (marked in blue) appeared in a small fraction of cells, thereby having relatively low clonal load, and therefore, their coexistence and contribution to the current clinical course is questionable. Such variants probably in-



■ ***Fig. 7.*** The clonal profiling of the variants detected in particular patients: *a* — a probe with five mutations in fore clusters; *b* — another probe with eight mutations in six clusters

dicate the possibility of new subclones genesis, and the diversification of a further clonal population within the existing tumor.

Matrix of primary and secondary mutations (see Fig. 3) shows that the primary mutations mostly developed in *DICER1*, *TP53*, and *ASXL1* among the examined samples. It is in compliance with the well-known fact that *ASXL1* gene is frequently affected primarily in patients with MDS [23]. A clonal evolution path with *TP53* as a driver mutation was also described [24]. *RUNX1* and *ASXL1* lesions were the most frequent among the secondary mutations.

Matrix of variants co-occurring in the same clusters (see Fig. 4) illustrates the lack of coexistence of mutations in genes of common biological function: different splicing genes (*SRSF2*, *SF3B1*), immune response genes (*LAG3*, *PDCD1*, *CD274*), which corresponds to the common trend of negative cooperativity where functionally similar genes are rarely affected together [25]. However, we have identified inconsistencies of such tendency: mutations in genes of microRNA processing (*DICER1*, *DROSHA*) appeared in the same clusters. Additionally, it is true for genes of epigenetic regulation (*TET2*, *ASXL1*, *EZH2*, *DNMT3A*, and *IDH1*), that was also observed in another research [23]. Such distinguishing tendencies might be related to the interdependent participation of these genes in the same biological pathway. It should be noted that the cluster variants with the lowest tumor burden (up to 10%) and duplicated probes were omitted from the resulting matrices, since their mutational load is comparable with the size of a method error [16].

One of the main potential applications of the tumor subclones identification and characterization is the monitoring of their size and genetic features in dynamics. Fig. 5 demonstrates an example of a negative tendency for disease progression where the growth of a primary clone and its subclone led to the emergence of another subclone.

For more reliable identification and subsequent exclusion of irrelevant polymorphisms from the list of detected variants, matching control material can be sequenced for each patient using tissue with a germinal genotype. This could also improve the copy-number variations calling as both tumor and normal genotype files can be introduced to CNVkit.

In some cases the tumor load can be evaluated by tissue morphology (in case of solid tumors) or flow cytometry: this would be beneficial to avoid the assumptions about the maximum allele frequency being indicative which were included when determining the tumor load using only NGS data: we hypothesized this indicative variant is heterozygous and not affected by indels (in the absence of obvious signs of the opposite).

■ ***Table 2***. Comparison of TS and WGS requirements

Platform	Cost (per sample, USD)	Depth	Data size (processed bam)
WGS	1000–3000	30–60	Depending on coverage ~60–350 GB
WES	500–2000	150–200	Depending on coverage ~5–20 GB
TS	300–1000	200–1000	Varies by panel size and coverage ~0.1–5 GB

The importance of subclones in tumor development has also been proven in other studies. In case of ovarian cancer heterogeneous clonal profiles of metastases led to different courses of tumor progression in each metastasis, which caused poor response to immunotherapy, limiting treatment options [26]. Clonal structure in MDS was also described earlier [1], but both papers imply the methods yet unavailable in clinical practice. In another study, clonal architecture of prostate cancer was reconstructed by different methods, and PyClone was shown to be appropriate in this case [27]. For their turn, we have demonstrated the suitability of computational methods in MDS, endorsing this technique and its beneficial prognostic value into clinical practice.

PyClone uses beta-binomial distribution to model mutation frequencies. There are alternatives: PhyloWGS [28] and DPClust [29] based on binomial distribution; but these tools were shown to be not applicable in all cases [27]: DPClust failed with the rate of 3.1% due to excessive computational memory (more than 250 GB), and PhyloWGS demanded inordinate runtime (more than 3 months). On the contrary, PyClone successfully completes all samples in the same study, which was confirmed by our work.

Moreover, the majority of these tools request data from the whole genome or exome sequencing (WGS, WES), which greatly exceeds the storage and computational resources needed for the same analysis performed with TS (Table 2 [30]).

In conclusion, the analysis of the clonal structure, being a modern trend in predicting the outcomes of tumors, is of highest importance for the prognosis and the selection of treatment, which establishes the relevance for pipeline development in case of MDS and other clonal diseases.

The authors have no conflicts of interest.

Financial support

This work was supported by the Russian Science Foundation under the grant No. 17-75-20145-П from 03.08.2020 "Search for potential therapeutic targets and biological prognostic factors in patients with myelodysplastic syndrome".

References

1. Nagata Y., Makishima H., Kerr C. M., Przychodzen B. P., Aly M., Goyal A., Awada H., Asad M. F., Kuzmanovic T., Suzuki H., Yoshizato T., Yoshida K., Chiba K., Tanaka H., Shiraishi Y., Miyano S., Mukherjee S., LaFramboise T., Nazha A., Sekeres M. A., Radivoyevitch T., Haferlach T., Ogawa S., Maciejewski J. P. Invariant patterns of clonal succession determine specific clinical features of myelodysplastic syndromes. *Nature Communications*, 2019, vol. 10, no. 1, p. 5386. doi:10.1038/s41467-019-13001-y
2. Montalban-Bravo G., Takahashi K., Patel K., Wang F., Xingzhi S., Nogueras G. M., Huang X., Pierola A. A., Jabbour E., Colla S., Gañan-Gomez I., Borthakur G., Daver N., Estrov Z., Kadia T., Pemmaraju N., Ravandi F., Bueso-Ramos C., Chamseddine A., Konopleva M., Zhang J., Kantarjian H., Futreal A., Garcia-Manero G. Impact of the number of mutations in survival and response outcomes to hypomethylating agents in patients with myelodysplastic syndromes or myelodysplastic/myeloproliferative neoplasms. *Oncotarget*, 2018, vol. 9, no. 11, pp. 9714–9727. doi:10.18632/oncotarget.23882
3. Landau D. A., Carter S. L., Stojanov P., McKenna A., Stevenson K., Lawrence M. S., Sougnez C., Stewart C., Sivachenko A., Wang L., Wan Y., Zhang W., Shukla S. A., Vartanov A., Fernandes S. M., Saksena G., Cibulskis K., Tesar B., Gabriel S., Hacohen N., Meyerson M., Lander E. S., Neuberg D., Brown J. R., Getz G., Wu C. J. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 2013, vol. 152, no. 4, pp. 714–726. doi:10.1016/j.cell.2013.01.019
4. Andor N., Graham T. A., Jansen M., Xia L. C., Aktipis C. A., Petritsch C., Ji H. P., Maley C. C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 2016, vol. 22, no. 1, pp. 105–113. doi:10.1038/nm.3984
5. Morris L. G. T., Riaz N., Desrichard A., Şenbabaoğlu Y., Hakimi A. A., Makarov V., Reis-Filho J. S., Chan T. A. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 2016, vol. 7, no. 9, pp. 10051–10063. doi:10.18632/oncotarget.7067
6. Zhang J., Fujimoto J., Zhang J., Wedge D. C., Song X., Zhang J., Seth S., Chow C.-W., Cao Y., Gumbs C., Gold K. A., Kalhor N., Little L., Mahadeshwar H., Moran C., Protopopov A., Sun H., Tang J., Wu X., Ye Y., William W. N., Lee J. J., Heymach J. V., Hong W. K., Swisher S., Wistuba I. I., Futreal P. A. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 2014, vol. 346, no. 6206, pp. 256–259. doi:10.1126/science.1256930
7. Tsvetkov N. Yu., Morozova E. V., Barkhatov I. M., Moiseev I. S., Barabanshchikova M. V., Tishkov A. V., Bug D. S., Petukhova N. V., Izmailova E. A., Bondarenko S. N., Afanasyev B. V. Prognostic value of next-generation sequencing data in patients with myelodysplastic syndrome. *Clinical Oncohematology*, 2020, vol. 13, no. 2, pp. 170–175 (In Russian). doi:10.21320/2500-2139-2020-13-2-170-175
8. Bolger A. M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, vol. 30, no. 15, pp. 2114–2120. doi:10.1093/bioinformatics/btu170
9. Poplin R., Ruano-Rubio V., DePristo M. A., Fennell T. J., Carneiro M. O., Van der Auwera G. A., Kling D. E., Gauthier L. D., Levy-Moonshine A., Roazen D., Shakir K., Thibault J., Chandran S., Whelan C., Lek M., Gabriel S., Daly M. J., Neale B., MacArthur D. G., Banks E. Scaling accurate genetic variant discovery to tens of thousands of samples. *Genomics*, 2017. doi:10.1101/201178
10. Li H., Durbin R. Fast and accurate short read alignment with Burrows — Wheeler transform. *Bioinformatics*, 2009, vol. 25, no. 14, pp. 1754–1760. doi:10.1093/bioinformatics/btp324
11. Benjamin D., Sato T., Cibulskis K., Getz G., Stewart C., Lichtenstein L. Calling somatic snvs and indels with mutect2. *Bioinformatics*, 2019. doi:10.1101/861054
12. McLaren W., Gil L., Hunt S. E., Riat H. S., Ritchie G. R. S., Thormann A., Flicek P., Cunningham F. The ensembl variant effect predictor. *Genome Biology*, 2016, vol. 17, no. 1, p. 122. doi:10.1186/s13059-016-0974-4
13. Wang K., Li M., Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 2010, vol. 38, no. 16, pp. e164–e164. doi:10.1093/nar/gkq603
14. Tange O. Gnu parallel-the command-line power tool. *The USENIX Magazine*, 2011, vol. 36, no. 1, pp. 42–47.
15. Genome Aggregation Database Consortium. Karczewski K. J., Francioli L. C., Tiao G., Cummings B. B., Alföldi J., Wang Q., Collins R. L., Laricchia K. M., Ganna A., Birnbaum D. P., Gauthier L. D., Brand H., Solomonson M., Watts N. A., Rhodes D., Singer-Berk M., England E. M., Seaby E. G., Kosmicki J. A., Walters R. K., Tashman K., Farjoun Y., Banks E., Poterba T., Wang A., Seed C., Whiffin N., Chong J. X., Samocha K. E., Pierce-Hoffman E., Zappala Z.,

O'Donnell-Luria A. H., Minikel E. V., Weisburd B., Lek M., Ware J. S., Vittal C., Armean I. M., Bergelson L., Cibulskis K., Connolly K. M., Covarrubias M., Donnelly S., Ferriera S., Gabriel S., Gentry J., Gupta N., Jeandet T., Kaplan D., Llanwarne C., Munshi R., Novod S., Petrillo N., Roazen D., Ruano-Rubio V., Saltzman A., Schleicher M., Soto J., Tibbetts K., Tolonen C., Wade G., Talkowski M. E., Neale B. M., Daly M. J., MacArthur D. G. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 2020, vol. 581, no. 7809, pp. 434–443. doi:10.1038/s41586-020-2308-7

16. Judkins T., Leclair B., Bowles K., Gutin N., Trost J., McCulloch J., Bhatnagar S., Murray A., Craft J., Wardell B., Bastian M., Mitchell J., Chen J., Tran T., Williams D., Potter J., Jammulapati S., Perry M., Morris B., Roa B., Timms K. Development and analytical validation of a 25-gene next generation sequencing panel that includes the BRCA1 and BRCA2 genes to assess hereditary cancer risk. *BMC Cancer*, 2015, vol. 15, no. 1, p. 215. doi:10.1186/s12885-015-1224-y
17. McCall C. M., Mosier S., Thiess M., Debeljak M., Pallavajjala A., Beierl K., Deak K. L., Datto M. B., Gocke C. D., Lin M.-T., Eshleman J. R. False positives in multiplex pcr-based next-generation sequencing have unique signatures. *The Journal of Molecular Diagnostics*, 2014, vol. 16, no. 5, pp. 541–549. doi:10.1016/j.jmoldx.2014.06.001
18. Poplin R., Chang P.-C., Alexander D., Schwartz S., Colthurst T., Ku A., Newburger D., Dijamco J., Nguyen N., Afshar P. T., Gross S. S., Dorfman L., McLean C. Y., DePristo M. A. A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 2018, vol. 36, no. 10, pp. 983–987. doi:10.1038/nbt.4235
19. Talevich E., Shain A. H., Botton T., Bastian B. C. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS Computational Biology*, 2016, vol. 12, no. 4, p. e1004873. doi:10.1371/journal.pcbi.1004873
20. Soong D., Stratford J., Avet-Loiseau H., Bahlis N., Davies F., Dispenzieri A., Sasser A. K., Schecter J. M., Qi M., Brown C., Jones W., Keats J. J., Auclair D., Chiu C., Powers J., Schaffer M. CNV Radar: an improved method for somatic copy number alteration characterization in oncology. *BMC Bioinformatics*, 2020, vol. 21, no. 1, p. 98. doi:10.1186/s12859-020-3397-x
21. Roth A., Khattra J., Yap D., Wan A., Laks E., Biele J., Ha G., Aparicio S., Bouchard-Côté A., Shah S. P. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 2014, vol. 11, no. 4, pp. 396–398. doi:10.1038/nmeth.2883
22. Miller C. A., McMichael J., Dang H. X., Maher C. A., Ding L., Ley T. J., Mardis E. R., Wilson R. K. Visualizing tumor evolution with the fishplot package for R. *BMC Genomics*, 2016, vol. 17, no. 1, p. 880. doi:10.1186/s12864-016-3195-z
23. Li X., Xu F., Wu L.-Y., Zhao Y.-S., Guo J., He Q., Zhang Z., Chang C.-K., Wu D. A genetic development route analysis on MDS subset carrying initial epigenetic gene mutations. *Scientific Reports*, 2020, vol. 10, no. 1, p. 826. doi:10.1038/s41598-019-55540-w
24. Chen J., Kao Y.-R., Sun D., Todorova T. I., Reynolds D., Narayanagari S.-R., Montagna C., Will B., Verma A., Steidl U. Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level. *Nature Medicine*, 2019, vol. 25, no. 1, pp. 103–110. doi:10.1038/s41591-018-0267-4
25. Sperling A. S., Gibson C. J., Ebert B. L. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nature Reviews Cancer*, 2017, vol. 17, no. 1, pp. 5–19. doi:10.1038/nrc.2016.112
26. Jiménez-Sánchez A., Memon D., Pourpe S., Veeraraghavan H., Li Y., Vargas H. A., Gill M. B., Park K. J., Zivanovic O., Konner J., Ricca J., Zamarin D., Walther T., Aghajanian C., Wolchok J. D., Sala E., Merghoub T., Snyder A., Miller M. L. Heterogeneous tumor-immune microenvironments among differentially growing metastases in an ovarian cancer patient. *Cell*, 2017, vol. 170, no. 5, pp. 927–938.e20. doi:10.1016/j.cell.2017.07.025
27. Liu L. Y., Bhandari V., Salcedo A., Espiritu S. M. G., Morris Q. D., Kislinger T., Boutros P. C. Quantifying the influence of mutation detection on tumour subclonal reconstruction. *Cancer Biology*, 2018. doi:10.1101/418780
28. Deshwar A. G., Vembu S., Yung C. K., Jang G. H., Stein L., Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 2015, vol. 16, no. 1, p. 35, doi:10.1186/s13059-015-0602-8
29. Nik-Zainal S., Van Loo P., Wedge D. C., Alexandrov L. B., Greenman C. D., Lau K. W., Raine K., Jones D., Marshall J., Ramakrishna M., Shlien A., Cooke S. L., Hinton J., Menzies A., Stebbings L. A., Leroy C., Jia M., Rance R., Mudie L. J., Gamble S. J., Stephens P. J., McLaren S., Tarpey P. S., Papaemmanuil E., Davies H. R., Varela I., McBride D. J., Bignell G. R., Leung K., Butler A. P., Teague J. W., Martin S., Jönsson G., Mariani O., Boyault S., Miron P., Fatima A., Langerød A., Aparicio S. A. J. R., Tutt A., Sieuwerts A. M., Borg Å., Thomas G., Salomon A. V., Richardson A. L., Børresen-Dale A.-L., Futreal P. A., Stratton M. R., Campbell P. J. The life history of 21 breast cancers. *Cell*, 2012, vol. 149, no. 5, pp. 994–1007. doi:10.1016/j.cell.2012.04.023
30. Bewicke-Copley F., Arjun Kumar E., Palladino G., Korfi K., Wang J. Applications and analysis of targeted genomic sequencing in cancer studies. *Computational and Structural Biotechnology Journal*, 2019, vol. 17, pp. 1348–1359. doi:10.1016/j.csbj.2019.10.004

УДК 57.087.1::616-006.4
doi:10.31799/1684-8853-2020-6-50-59

Построение и апробация биоинформатического пайплайна для определения клональных профилей при миелодиспластическом синдроме

Д. С. Буг[а], специалист, orcid.org/0000-0002-5849-1311, dmitriybs@1spbgmu.ru
А. А. Приходько[а], студентка, orcid.org/0000-0002-0001-7932
Е. А. Бакин[а,б], канд. техн. наук, доцент, orcid.org/0000-0002-5694-4348
А. В. Тишков[а], канд. физ.-мат. наук, доцент, orcid.org/0000-0002-4282-8717
Н. В. Петухова[а], канд. биол. наук, orcid.org/0000-0001-6397-824X
И. М. Бархатов[а], канд. мед. наук, orcid.org/0000-0002-8000-3652
Е. В. Морозова[а], канд. мед. наук, доцент, orcid.org/0000-0002-9605-485X
И. С. Моисеев[а], доктор мед. наук, доцент, orcid.org/0000-0002-4332-0114
[а]Первый Санкт-Петербургский государственный медицинский университет им. акад. И. П. Павлова, Льва Толстого ул., 6-8, Санкт-Петербург, 197022, РФ
[б]Санкт-Петербургский государственный университет аэрокосмического приборостроения, Б. Морская ул., 67, Санкт-Петербург, 190000, РФ

Введение: результаты последних научных исследований показывают более очевидной связь между клональным профилем опухоли и его клиническим значением. Однако в настоящее время отсутствует доступный в клинической практике и надежный метод для клонального профилирования. Миелодиспластический синдром представляет собой сложную клональную патологию гемопоэтической стволовой клетки, характеризующуюся морфологической дисплазией, цитопенией и высоким риском трансформации в острый миелоидный лейкоз. Клинический исход миелодиспластического синдрома может быть чрезвычайно гетерогенным, поэтому для выяснения истинного состояния гемопоэтической системы и дальнейшего прогноза больных с такими сложными нарушениями необходимо специфическое изучение клональных профилей. **Цель исследования:** разработка пайплайна, предназначенного для определения клональных профилей пациентов с миелодиспластическим синдромом на основе данных таргетного секвенирования следующего поколения. **Результаты:** пайплайн был разработан и апробирован на выборке из 35 пациентов с миелодиспластическим синдромом преимущественно высокого риска. Показана возможность использовать данные таргетного секвенирования для оценки гетерогенности клональных профилей и характеристики их генных свойств. Данный подход позволит идентифицировать соответствие между типом индивидуального профиля и прогнозом заболевания, влиять на выбор терапии. Продемонстрированы характеристики полученных клональных профилей и описан процесс их анализа. **Практическая значимость:** информация о выявленных закономерностях и взаимосвязи между характеристиками клонального профиля (количеством субклонов, частотой мутаций на клон) и клиническим исходом может быть использована врачами в современной практике для более точного подбора терапии в зависимости от выявленной индивидуальной специфичности заболевания.

Ключевые слова — миелодиспластический синдром, биоинформатический пайплайн, клональный профиль, первичная мутация, субклон, таргетное секвенирование, секвенирование следующего поколения.

Для цитирования: Bug D. S., Prikhodko A. A., Bakin E. A., Tishkov A. V., Petukhova N. V., Barkhatov I. M., Morozova E. V., Moiseev I. S. Building and evaluation of bioinformatic pipeline for determination of clonal profiles in myelodysplastic syndrome. *Информационно-управляющие системы*, 2020, № 6, с. 50–59. doi:10.31799/1684-8853-2020-6-50-59
For citation: Bug D. S., Prikhodko A. A., Bakin E. A., Tishkov A. V., Petukhova N. V., Barkhatov I. M., Morozova E. V., Moiseev I. S. Building and evaluation of bioinformatic pipeline for determination of clonal profiles in myelodysplastic syndrome. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2020, no. 6, pp. 50–59. doi:10.31799/1684-8853-2020-6-50-59