"Somatic" and "pathogenic" - is the classification strategy applicable in times of large-scale sequencing?

Constance Baer, Wencke Walter, Stephan Hutter, Sven Twardziok, Manja Meggendorfer, Wolfgang Kern, Torsten Haferlach and Claudia Haferlach

MLL Munich Leukemia Laboratory, Munich, Germany

Ferrata Storti Foundation

Introduction

In the early days of sequencing only a small number of bases was evaluated because of the labor-intensive nature of the procedure. Genes were identified to play a role in the pathogenesis of neoplasms in animal models and cell lines. Subsequently, these mutations were analyzed in samples from patients and their impact on prognosis was evaluated. The list of examples is long: e.g. *TP53* was found to be universally mutated across cancers,[1] and *NPM1* is now among the most frequently analyzed genes in acute myeloid leukemia[2] and the mutation defines its own acute myeloid leukemia subtype in the current World Health Organization classification.[3]

High-throughput sequencing has changed the landscape. It is now possible to sequence a huge number of genes up to exomes and even whole genomes in a comparably short time at affordable cost. The challenge is no longer the sequencing, but rather the evaluation of the results and the interpretation of their impact on diagnosis, prognosis or therapeutic decisions. This has led to some major changes in the way we view sequencing data. In 2015, the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) recommended changing the terms mutation and polymorphism to "*variant*". Variants are then further subdivided into five categories depending on the likelihood of their association with the disease.[4] The definition was designed for hereditary diseases and therefore addresses germline variants.

The vast majority of genetic events in cancer are somatic.[5] Acquired variants represent potential drug targets or biomarkers. Testing a sample, e.g. of colon cancer, for mutations is frequently performed by comparing variants from a biopsy to those in leukocytes as reference to identify only somatic variants (tumor-normal comparison).[6,7] The classical tumor-normal workflow is challenging in studies with large or historic cohorts because of additional sequencing costs, or limited availability of reference material. Leukemia represents another challenge, since blood cannot be used as the reference material. In addition, the growing knowledge of genetic complexity and tumor heterogeneity challenges the historic binary variant classifications (mutation, polymorphism) in the somatic field as well (Figure 1).

Today

Ideally, the results of tumor sequencing are compared to those of reference material with an unaltered germline sequence. Clonal hematopoiesis of indeterminate potential has made us familiar with the idea that mutations are acquired as part of the aging process.[8] Blood cells are strongly affected by the continuous accumulation of somatic changes as a consequence of lifelong proliferation,[9] but the phenomenon could apply to all types of reference material.[10,11] Tissue formed of cells that divide less quickly (e.g. cerebral tissue)[12] would be preferred as a reference; however, this is not a practical approach for routine analysis. Easily accessible sources of reference material are hair follicles, nails, urine, T cells, fibroblast cultures, buccal swabs, saliva, and skin biopsies, but poor DNA yield and the presence of leukemic cell contamination (e.g. DNA from blood in nails) are potential challenges to the use of such material.[13-15]

In the absence of available reference material, the variant allele frequency (VAF) can be used to distinguish germline from somatic variants. A germline variant is present with a 50% (heterozygous) or 100% (homozygous) VAF. An acquired variant is usually present with a lower VAF, because it is not present in all cells. The caveat is that other factors can also contribute to VAF. Firstly, technical issues

(polymerase chain reaction/amplification bias) can contribute to skewed VAF. Secondly, somatic mutations can also occur with a VAF of 50% if the number of malignant cells in the analyzed sample is high. Lastly, genetic features influence the VAF. A deletion of 17p, would cause all germline variants in the deleted locus (foremost *TP53*) to appear with VAF not around 50%. Copy number gains and copy neutral loss of heterozygosity also influence VAF.

Databases are frequently used for variant interpretation. Curated databases for variants meet high quality standards, but contain only a small number of variants.[16] Databases with large numbers of contributors, such as dbSNP, contain more variants. Initially, many variants were classified as benign germline variants if they were listed in the dbSNP.[17] All databases, which allow individual submitters to add data, are error prone. The diversity of the exome was impressively demonstrated by Lek *et al.*, who analyzed 60,706 individuals of different ethnicities and found on average one variant for every eight bases of the exome. Data were collected by the Exome Aggregation Consortium (ExAC), which included a col-

lection of whole exome sequencing data from a broad range of studies. ExAC has recently been extended to genomes (gnomAD).[18] As of today, gnomAD contains data from 141,456 individuals. This allows for a representative overview of the population and ethnic groups. As such, it can be used to exclude frequent variants in the population from candidate disease-associated aberrations. However, it must be kept in mind that the metadata from such a collection includes many blood samples and even individuals with an (undiagnosed) hematologic disease. The V617F mutation in the *JAK2* gene, which is found in the majority of patients with myeloproliferative neoplasms, is also found in 0.04% of individuals in the gnomAD dataset, so variants in gnomAD cannot be classified *per se* as benign.

Considering the difficulties, it can be questioned whether the strict separation of somatic and germline is necessary. There are a number of reasons why they should be distinguished: (i) only somatic changes prove clonal outgrowth in the case of clonal hematopoiesis of indeterminate potential and clonal cytopenia of undetermined significance[8] and they are used for the calculation
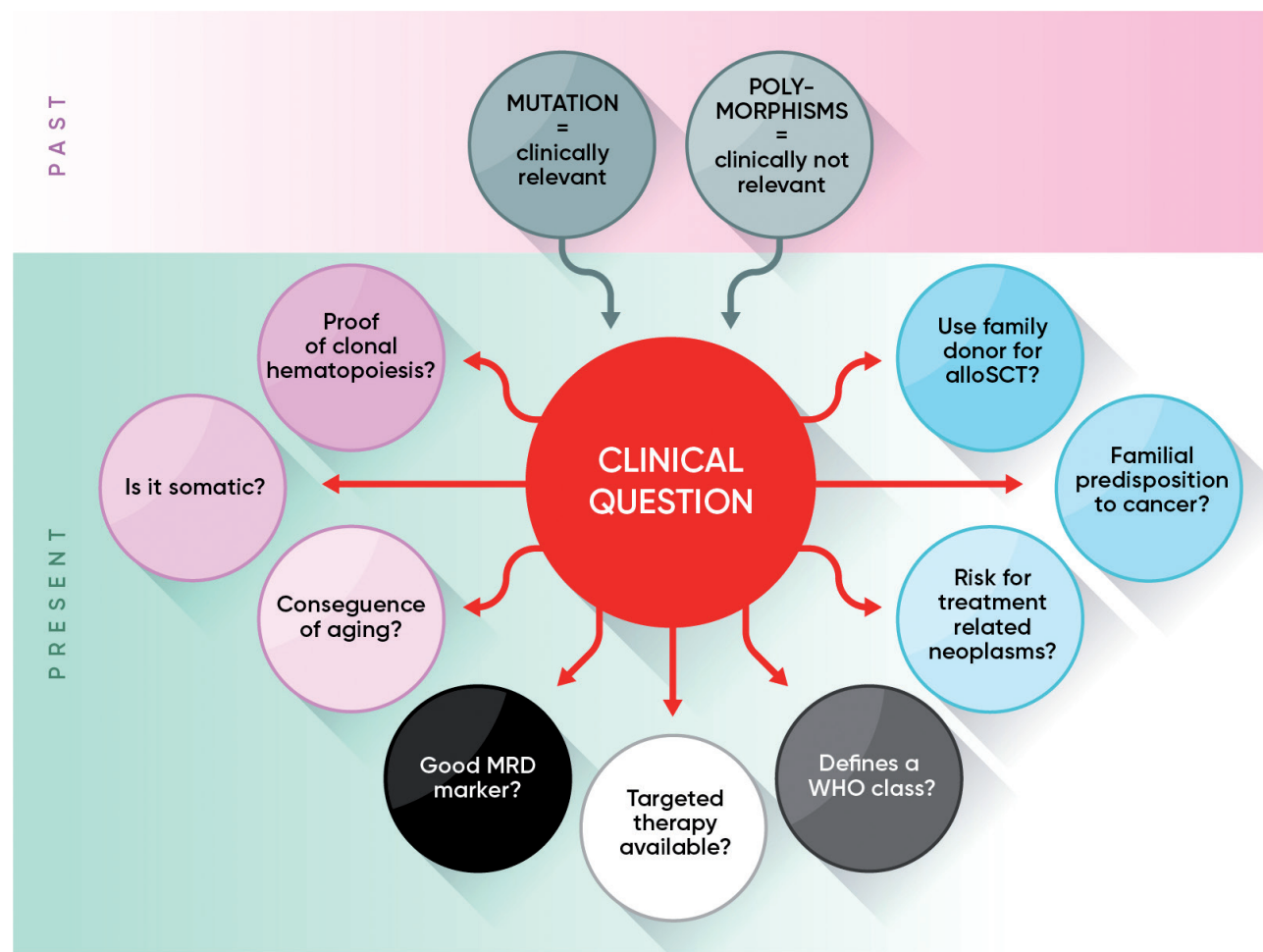
Figure 1. Clinical questions for current variant classifications. The complexity of variant classification challenges the terms "mutation" and "polymorphism", which were clearly associated with clinical relevance in the past. The classification, interpretation and consequences of each variant depend on the context. The identification of somatic changes proves clonality, and somatic aberrations qualify as markers of minimal/measurable residual disease. Other clinical issues are the availability of targeted therapies for mutations and the World Health Organization classification, which genetically defines certain entities. Finally, inherited variants are not only discussed in the context of familial predisposition, but are also relevant if family members are considered as stem cell donors. AlloSCT: allogeneic stem cell transplantation; MRD: minimal/measurable residual disease; WHO: World Health Organization.

of tumor mutation burden;[19] (ii) the presence of a genetic aberration is frequently used as a marker for monitoring minimal (measurable) residual disease (MRD).[20] A germline variant would not be eradicated by treatment (with the exception of allogeneic bone marrow transplantation), and is therefore not informative as a MRD marker; (iii) germline variants can predispose to or cause cancer and other diseases. Tarailo-Graovac *et al.* found that 2.8% of individuals from the ExAC database have variants which are implicated in a wide variety of Mendelian disorders.[21] The 2017 World Health Organization classification now also recognizes neoplasms with germline predisposition and mutations in *ETV6*, *RUNX1* and other genes.[3] Knowledge of pathogenic germline variants is of importance for family counseling and to determine whether family members can be considered as stem cell donors. Germline variants are important modulators of a patient's outcome and treatment response. An aberration in the DNA damage repair pathway could increase the risk of therapy-related myeloid neoplasms.[22]

An impressive example that the variant itself is of greater importance than its origin is the response to treatment with olaparib. Patients respond if they have either a germline or acquired variant of *BRCA1/2*.[23] It is therefore essential to know whether the mutation is pathogenic or actionable. The definition of pathogenic or actionable is not trivial and is highly context-dependent. For inherited diseases, a pathogenic variant is usually understood as being causal. A clear variant-phenotype relationship has been recognized in a few cases (e.g. cystic fibrosis), but for many other disease types such a relationship is more elusive. The five-tier system for the classification of hereditary variants currently recommended by the ACMG and AMP also recognizes the categories of "likely pathogenic" and "likely benign".[4]

Here, we focus on variants in cancer. When translating genetic findings into the clinic, a variant might have a different value depending on the immediate question at hand. Sukhai *et al.* proposed the term "actionable" to describe variants which affect patients' management.[24] The 2017 guidelines for variants in cancer from the AMP, American Society of Clinical Oncology (ASCO), and the College of American Pathologists (CAP) suggest a four-tier system: tier I, strong clinical significance; tier II, potential clinical significance; tier III, unknown clinical significance; and tier IV, a benign or likely benign variant (Figure 2).[25] The interpretation should be further subdivided into the categories "*diagnostic*", "*prognostic*" and "*predictive*". It is suggested that the four-tier system is applied to each category. The highest level of evidence is given to biomarkers that predict response or resistance to Food and Drug Administration-approved therapies and to variants from professional diagnostic and prognostic guidelines.[25] The *BRAF* V600E mutation is such an example, as its presence allows vemurafenib treatment.[26]

At first glance, a tiered system seems to be a well-standardized approach, but each case presents its own challenges. One example would be a patient with cytopenia
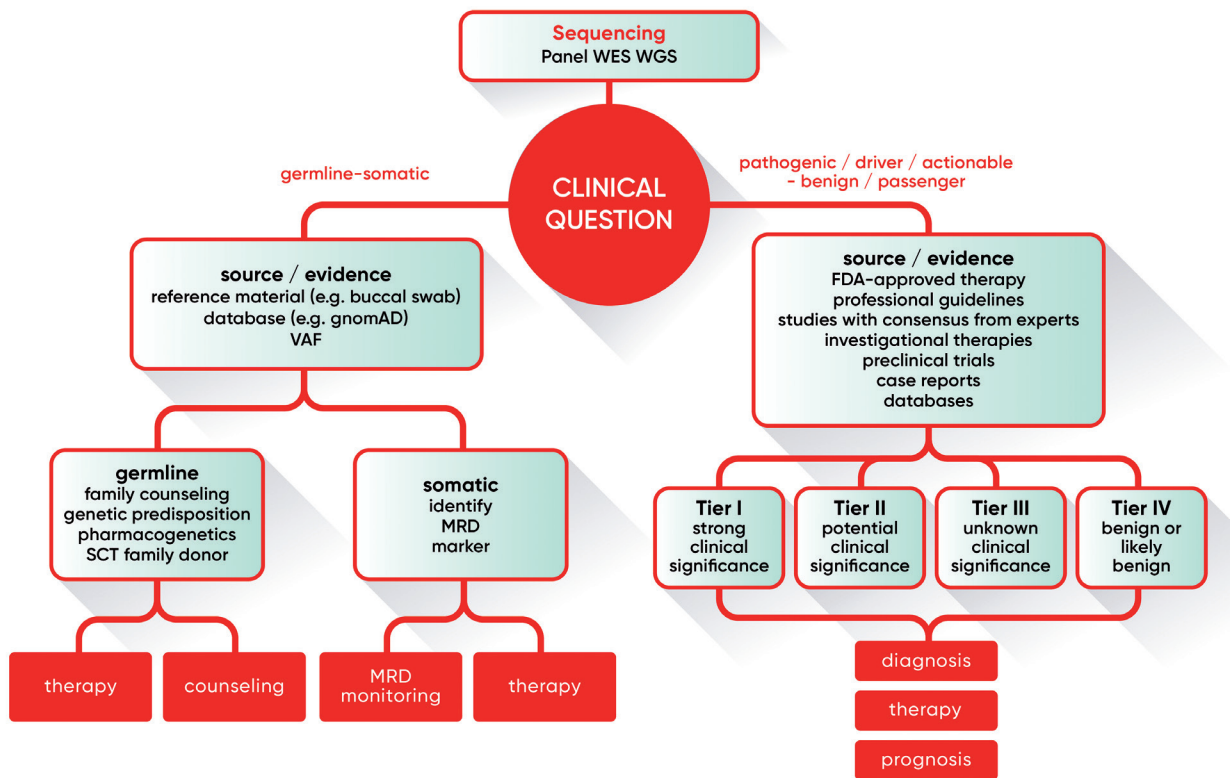
Figure 2. Decision tree for variant classification and clinical-decision making. Sequencing data are used to answer different clinical questions. Here we separate the issue of germline *versus* somatic and biological functions (e.g. pathogenic *vs.* benign). The questions are closely related in everyday life, however the sources and evidence supporting decisions are different (examples are outlined here). The definitions of the clinical significance of the four-tier classification system are from the guidelines by Li *et al*.[25] WES: whole-exome sequencing; WGS: whole-genome sequencing; VAF: variant allele frequency; FDA: Food and Drug Administration; SCT: stem cell transplantation; MRD: minimal/measurable residual disease.

and an acquired variant in the *TET2* gene. The variant could certainly prove clonality, which is essential for diagnosing clonal cytopenia of undetermined significance. The variant would therefore be tier I or tier II in the diagnostic classification. However, if the same variant is found in an patient with acute myeloid leukemia, for whom the therapeutic procedure remains to be defined, the classification becomes more complicated. Currently, no specific therapy is available for *TET2*-positive malignancies. The variant could therefore end up in tier III if a strict interpretation of the guidelines is used.

Hematologic diseases are closely related and mutations are generally typical, but not exclusive to one disease. For example, the L265P variant in *MYD88* is found in 90% of all patients with Waldenström macroglobulinemia, but is also present, at a lower frequency, in patients with other B-cell neoplasms.[27] Consequently, no variant has the sensitivity or specificity to qualify as "diagnostic" when the rules are applied strictly. For a patient, all three of the categories (*diagnostic, prognostic, predictive*) are important. The presence of a *MYD88* mutation suggests that ibrutinib therapy is an option.[28] Therefore, L265P would probably be classified as tier I in most interpretations.

Unlike *MYD88* L265P, many genes do not have well-described mutation hotspots. A large variety of variants are observed in genes such as *RUNX1*, *CEBPA* and *DNMT3A*. Databases can be helpful when the variant has been described before. An example is the Catalog of Somatic Mutations in Cancer (COSMIC),[29] which has collected information on mutations from peer-reviewed journals since 2004.[30] The UMD database contains 6,870 variants for *TP53* alone.[31] COSMIC is manually curated and the UMD-TP53 has developed its own data-driven curation strategy.[32] For genes which are less well understood than *TP53*, databases are less helpful.

Alternatively, algorithms that predict the effect of variants on protein structure could be used. *In silico* analyses are immediately available and can be performed without expert knowledge. Major influencing factors include the type of amino acid exchange and the location of the variant in conserved or functional domains. The dbNSFP database[33] contains pre-calculated values for all possible single nucleotide variants which result in amino-acid or splice-site changes in the human genome from 18 different algorithms. However, the read-out is not necessarily "yes" or "no". Different algorithms almost never come to exactly the same result. Some well-known pathogenic variants are not rated high enough by common algorithms. For example, the W515L mutation in the *MPL* gene is known to be typical in myeloproliferative neoplasms, but is rated as damaging/pathogenic by only seven of the 18 algorithms. Currently, most of the algorithms are trained on single nucleotide variants and cannot be applied to indels. Finally, it should be emphasized that a high pathogenicity score is not always a synonym for causality or actionability.

Tomorrow

There is no one-step solution for variant classification. Searching software for genetic variant interpretation on Google gives millions of hits. In a cross-laboratory comparison of variant classification, there was only 34% concordance.[34] The data provided by whole exome sequencing or whole genome sequencing will inevitably make the analysis more complex, since the number of identified variants is many times greater than with panel sequencing. Variants will be found in all the approximately 20,000 human genes, but not all of them will be either relevant or redundantly mutated in a disease. A comparison of two large study sets (TCGA and BeatAML)[35] confirmed that 33 genes are frequently mutated in acute myeloid leukemia, but there was diversity regarding genes with a mutation frequency of 2% or less. Over 2,000 genes were found to be mutated in only one of the datasets or even one patient. Current guidelines, such as those from the AMP,[36] are based on characterized genes. Therefore, Kaur *et al.*[7] argue that panels are preferred for breast cancer.[37] They can achieve deeper coverage, which is synonymous with greater sensitivity. Sensitivity in the 1-3% range is increasingly required because subclones should be detected,[38] or because clonal hematopoiesis of indeterminate potential is already diagnosed if a mutation is found with a VAF of 2%.[39] The sensitivity of whole genome sequencing is currently in the range of 15-20%, but the technique allows simultaneous identification of structural and copy number variations and the cost of sequencing is decreasing.[40]

We therefore suggest short-term strategies for current diagnostic use of large-scale datasets and long-term approaches to advance our understanding of the malignant process and therapeutic options.

Overstating the importance of a variant is clearly dangerous, but a report with variants of unknown clinical significance can be difficult to translate into clinical consequences. Here, we outline major aspects for today's usage. First, collaboration between different laboratory branches is needed to compare genetic and other biomarkers and between laboratories and physicians to tailor personalized answers. For example, by integration of different laboratory results, a patient in remission according to morphology, but still with a VAF of 50% could be identified to have a rare and possibly less relevant germline variant. Another example derives from the growing awareness of germline predisposition e.g. with *SAMD9/SAMD9L* mutations in myelodysplastic syndromes.[41] If the family background and reference material are provided, testing can be adjusted. Second, databases are the cornerstones of variant interpretation. An impressive example of the success of combined forces is gnomAD, which is now the worldwide reference for germline variants. Third, in the context of monitoring, serial testing can reveal the outgrowth of a clone with a specific variant and highlight clinical relevance, as demonstrated by retrospective studies.[38,42] Well-documented information from multiple time points is a resource for variant classification, also for following patients with the same variant, and ideally should be included in databases and classification algorithms in the future. Finally, filtering for known and well-studied changes is always a valid first step. The first whole-genome sequencing studies in hematology demonstrated respectable sensitivity and specificity when filtering for known copy number variations, structural variations and genes.[43,44]

Mutations outside coding regions are difficult to associate with functions. They influence gene expression by altering transcription factor binding, alternative splicing, and certain genomic variants are likely to be causal for the acquisition of chromosomal aberrations.[45-47]

Furthermore, they can influence pharmacogenetics,[48] and the effect of the same somatic mutation may differ between patients depending on other acquired or inherited genetic factors.

Artificial intelligence is a logical choice to exploit the full potential of the data and leave the binary mutation/polymorphism classification behind. The use of artificial intelligence in clinical oncology, genome interpretation, and especially in variant reporting has gained momentum.[49,50] Data available from manually classified variants can be used to train deep neural networks.[49] The advantage of this approach is that the algorithm is able to autonomously extract relevant features for classification and identify important combinations not only for genetic information but for all types of biomarker. There is no need for any manually defined set of rules. This is especially useful for variant interpretation because, as described above, it is basically impossible to capture the entire complexity using a simple set of rules. The output of such algorithms could indicate clinically relevant likelihoods. However, in order to aid clinical decision-making, a future report should not just be a list of variants with their individual classifications but rather a personalized summary of all genetic information (including structural and copy number variations) and other biomarkers and their combined meaning.

References

1. Koshland DE Jr. Molecule of the year. Science. 1993;262(5142):1953.
2. Heath EM, Chan SM, Minden MD, et al. Biological and clinical consequences of NPM1 mutations in AML. Leukemia. 2017;31(4):798-807.
3. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127 (20):2391-2405.
4. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-424.
5. Vogelstein B, Papadopoulos N, Velculescu VE, et al. Cancer genome landscapes. Science. 2013;339(6127):1546-1558.
6. Mwenifumbo JC, Marra MA. Cancer genome-sequencing study design. Nat Rev Genet. 2013;14(5):321-332.
7. Doroshow JH, Kummar S. Translational research in oncology--10 years of progress and future prospects. Nat Rev Clin Oncol. 2014;11(11):649-662.
8. Bejar R. CHIP, ICUS, CCUS and other four-letter words. Leukemia. 2017;31(9):1869-1871.
9. Lee-Six H, Obro NF, Shepherd MS, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561(7724):473-478.
10. Fernandez LC, Torres M, Real FX. Somatic mosaicism: on the road to cancer. Nat Rev Cancer. 2016;16(1):43-55.
11. Yokoyama A, Kakiuchi N, Yoshizato T, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312-317.
12. Lynch M. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A. 2010;107(3):961-968.
13. Padron E, Ball MC, Teer JK, et al. Germ line tissues for optimal detection of somatic variants in myelodysplastic syndromes. Blood. 2018;131(21):2402-2405.
14. Thiede C, Prange-Krex G, Freiberg-Richter J, Bornhauser M, Ehninger G. Buccal swabs but not mouthwash samples can be used to obtain pretransplant DNA fingerprints from recipients of allogeneic bone marrow transplants. Bone Marrow Transplant. 2000;25 (5):575-577.
15. Imanishi D, Miyazaki Y, Yamasaki R, et al. Donor-derived DNA in fingernails among recipients of allogeneic hematopoietic stem-cell transplants. Blood. 2007;110(7):2231-2234.
16. Ritter DI, Roychowdhury S, Roy A, et al. Somatic cancer variant curation and harmonization through consensus minimum variant level data. Genome Med. 2016;8(1):117.
17. Kitts A, Phan L, Ward M, Holmes B. The Database of Short Genetic Variation (dbSNP). The NCBI Handbook, 2nd ed. National Center for Biotechnology Information (US), Bethesda (MD). 2013:259-298.
18. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616): 285-291.
19. Yarchoan M, Hopkins A, Jaffee EM. Tumor mutational burden and response rate to PD-1 inhibition. N Engl J Med. 2017;377 (25):2500-2501.
20. Jongen-Lavrencic M, Grob T, Hanekamp D, et al. Molecular minimal residual disease in acute myeloid leukemia. N Engl J Med. 2018;378(13):1189-1199.
21. Tarailo-Graovac M, Zhu JYA, Matthews A, van Karnebeek CDM, Wasserman WW. Assessment of the ExAC data set for the presence of individuals with pathogenic genotypes implicated in severe Mendelian pediatric disorders. Genet Med. 2017;19(12):1300-1308.
22. McNerney ME, Godley LA, Le Beau MM. Therapy-related myeloid neoplasms: when genetics and environment collide. Nat Rev Cancer. 2017;17(9):513-527.
23. George A, Banerjee S, Kaye S. Olaparib and somatic BRCA mutations. Oncotarget. 2017;8(27):43598-43599.
24. Sukhai MA, Craddock KJ, Thomas M, et al. A classification system for clinical relevance of somatic variants identified in molecular profiling of cancer. Genet Med. 2016;18 (2):128-136.
25. Li MM, Datto M, Duncavage EJ, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017;19(1):4-23.
26. Dietrich S, Glimm H, Andrulis M, et al. BRAF inhibition in refractory hairy-cell leukemia. N Engl J Med. 2012;366(21):2038-2040.
27. Yu X, Li W, Deng Q, et al. MYD88 L265P mutation in lymphoid malignancies. Cancer Res. 2018;78(10):2457-2462.
28. Treon SP, Xu L, Hunter Z. MYD88 Mutations and response to ibrutinib in Waldenstrom's macroglobulinemia. N Engl J Med. 2015;373(6):584-586.
29. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45(D1): D777-D783.
30. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91(2):355-358.
31. Soussi T, Wiman KG. TP53: an oncogene in disguise. Cell Death Differ. 2015;22(8):1239-1249.
32. Edlund K, Larsson O, Ameur A, et al. Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. Proc Natl Acad Sci U S A. 2012;109(24):9551-9556.
33. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat. 2016;37(3):235-241.
34. Amendola LM, Jarvik GP, Leo MC, et al. Performance of ACMG-AMP variant-interpretation guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet. 2016;98(6):1067-1076.
35. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562 (7728):526-531.
36. McClure RF, Ewalt MD, Crow J, et al. Clinical significance of DNA variants in chronic myeloid neoplasms: a report of the Association for Molecular Pathology. J Mol Diagn. 2018;20(6):717-737.
37. Kaur P, Porras TB, Ring A, Carpten JD, Lang JE. Comparison of TCGA and GENIE genomic datasets for the detection of clinically actionable alterations in breast cancer. Sci Rep. 2019;9(1):1482.
38. Chen J, Kao YR, Sun D, et al. Myelodysplastic syndrome progression to acute myeloid leukemia at the stem cell level. Nat Med. 2019;25(1):103-110.
39. Valent P, Kern W, Hoermann G, et al. Clonal

hematopoiesis with oncogenic potential (CHOP): separation from CHIP and roads to AML. Int J Mol Sci. 2019;20(3).

40. Wrzeszczynski KO, Felice V, Abhyankar A, et al. Analytical validation of clinical whole-genome and transcriptome sequencing of patient-derived tumors for reporting targetable variants in cancer. J Mol Diagn. 2018;20(6):822-835.

41. Kennedy AL, Shimamura A. Genetic predisposition to MDS: clinical features and clonal evolution. Blood. 2019;133(10):1071-1085.

42. da Silva-Coelho P, Kroeze LI, Yoshida K, et al. Clonal evolution in myelodysplastic syndromes. Nat Commun. 2017;8:15099.

43. Klintman J, Barmpouti K, Knight SJL, et al. Clinical-grade validation of whole genome sequencing reveals robust detection of low-frequency variants and copy number alterations in CLL. Br J Haematol. 2018;182(3):412-417.

44. Höllein A, Twardziok SO, Walter W, et al. WGS and RNA-seq is superior to conventional diagnostic tests in multiple myeloma: ready for prime time? Blood. 2018;132 (Suppl. 1):4442.

45. Khurana E, Fu Y, Chakravarty D, et al. Role of non-coding sequence variants in cancer. Nat Rev Genet. 2016;17(2):93-108.

46. Pellagatti A, Armstrong RN, Steeples V, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood. 2018;132(12):1225-1240.

47. Loh PR, Genovese G, Handsaker RE, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature. 2018;559(7714):350-355.

48. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature. 2015;526(7573):343-350.

49. Sundaram L, Gao H, Padigepati SR, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161-1170.

50. Zomnir MG, Lipkin L, Pacula M, et al. Artificial intelligence approach for variant reporting. JCO Clin Cancer Inform. 2018;2.