A decade with whole exome sequencing in haematology

Hansen, Marcus C; Haferlach, Torsten; Nyvold, Charlotte G

Published in:
British Journal of Haematology

DOI:
10.1111/bjh.16249

Publication date:
2020

Document version:
Accepted manuscript

MR. MARCUS CELIK HANSEN (Orcid ID : 0000-0003-3083-4850)

Article type : Reviews

A decade with whole exome sequencing in hematology

Marcus C. Hansen[1#], Torsten Haferlach[2], Charlotte G. Nyvold[1]

1) Odense University Hospital, Hematology Pathology Research Laboratory, Research Unit for Hematology and Research Unit for Pathology, University of Southern Denmark, Odense, DK

2) MLL Munich Leukemia Laboratory GmbH, Munich, DE

Corresponding author

Contact information:

Marcus C. Hansen

E-mail: marcus.celik.hansen@rsyd.dk

Summary

The first decade of capture-based targeted whole exome sequencing (WES) has now passed, while the sequencing modality continues to find more widespread usage in clinical research laboratories and still offers an unprecedented diagnostic assay in terms of throughput, informational content and running costs. Until quite recently, WES has been out of reach for many clinicians and molecular biologists, and it still poses issues or is met with some

reluctance with regards to cost versus benefit in terms of effective assay costs, hands-on laboratory work and data analysis bottlenecks. Although WES is used more than ever, it may also be argued that the usage is peaking and that new implementations, or relevance in its current state, will likely be leveling off during the following decade as the price on whole genome sequencing continues to drop. In this review, we focus on the past decade of targeted whole exome sequencing in malignant hematology. We thematically revisit some of the significant discoveries and niches that use next-generation sequencing, and we outline what and how WES has contributed to the field – from clonal hematopoiesis of the aging bone marrow to profiling malignancies down to the single cell.

Introduction

The human exome, which holds the protein coding information, constitutes one to two percent of the total genome. Errors in these regions of the genome are by far the largest contributors to Mendelian diseases (Chong, et al 2015) and acquired clonal diseases, such as cancer. Whole exome sequencing (WES) has been instrumental in elucidating the latter in the field of hematology and the complex landscapes of hematological malignancies.

Illustrative paraphrasing, as exemplified above, frequently finds usage in papers relating to WES but does not necessarily capture the contributions and shortcomings of this groundbreaking invention. Without much ado, the 10-year anniversary of capture-based targeted exome sequencing is now passing, while the technique continues to find more widespread usage in clinical research laboratories and still offers an unprecedented diagnostic assay in terms of throughput, informational content and running costs. Until quite recently, exome sequencing has been out of reach for many clinicians and molecular biologists, and it still poses issues or is met with some reluctance with regards to cost *versus* benefit in terms of effective assay costs, hands-on laboratory work and data analysis bottlenecks. In some situations, targeted panel or amplicon sequencing is more suitable, e.g., in focused clinical tests. Although WES is now used more than ever, it may also be argued that this sequencing modality is peaking and that new implementations, or usage in its current state, will likely be leveling off during the next decade.

In fact, *exomic sequencing* was not initiated in 2009, but several years earlier. Apart from the strategy of narrowing whole genome sequencing analyses to the coding parts, one interesting

approach was the daunting design of PCR primers targeting coding sequences of 18,191 genes in the study of the mutational *landscape* of colorectal and breast cancers (Wood*, et al* 2007). The jargon was further elaborated by emphasizing that candidate cancer driver mutations do not just appear as recurrent gene "mountains" but also as gene "hills" that are large in number. The latter whimsically described the long tail of low-frequent somatic mutations, which is nicely exemplified by chronic lymphocytic leukemia (CLL) in hematology. Looking back on the first decade of whole genome sequencing, Vogelstein and fellow researchers pondered the genomic landscapes of common cancers and the lessons learned from these studies (Vogelstein*, et al* 2013), with an estimated price of more than $100,000 per sequenced genome. First, the number of somatic mutations is somewhat representative of the specific cancer type, which is a notion still used today. Second, the number of mutations increases with age. Third, the estimated number of canonical mutated driver genes was at the time surprisingly small, with 125 defined based on the evaluation of more than three thousand tumors and a two orders of magnitude larger number of somatic mutations. Certainly, the "completion" of the genome in 2003 marks an important reference point in the early onset of the genomic revolution in cancer research. Naturally, sequencing of the human genome was a massive research undertaking that led to a public draft of the human genome by the International Human Genome Sequencing Consortium (Lander*, et al* 2001), made freely available in 2001 and finalized in 2003. In parallel, Venter *et al* published their version of the genome (Venter*, et al* 2001) and estimated that exons span 1.1% of the genome. One of the major differences between these studies was the initial estimation of the number of genes – an issue which remains unsettled.

In hindsight, exome sequencing has been of tremendous value, taught us much about the genome, and has shed light on the pathobiology of a wide range of diseases, especially cancers. The coding genome contains roughly twenty thousand single-nucleotide variants (Choi*, et al* 2009, Ng*, et al* 2009). Not only is this an inherent feature of the human genome, it is also a practical handle for quality assessment, as too high or low a number of coding single-nucleotide polymorphisms (SNPs) in a sample of a given population indicates a problem with sequencing, library or DNA sample. The typical or expected number of mutations of a specific cancer type and age constitutes another example of such an attribute. The number of mutations generally increases with age, and the combinations of these mutations occur in a

nonrandom fashion (Welch*, et al* 2012). If the number of coding variants corresponds to the number of genes (Choi*, et al* 2009, Ng*, et al* 2009), then it is highly interesting that, on average, one variant can be expected per gene, which is also supported by data submitted in relation to exome sequencing of colorectal cancer (Goryca*, et al* 2018). Furthermore, very few nonsense mutations are generally found by sequencing (Choi*, et al* 2009).

While exome sequencing has taken a relatively long time to find its way to the clinic, and for many laboratories, perhaps it will never enter the diagnostic scene; in parallel, many research studies have included massive undertakings of collectively gathering of samples from thousands of individuals (Genomes Project*, et al* 2015). The Exome Aggregation Consortium has managed to combine more than sixty thousand exomes (Lek*, et al* 2016) from 22 projects, with *1000 Genomes* and *The Cancer Genome Atlas* included. The *hyperauthorships* of the last two decades in life sciences are indicators of the sheer size of such projects and the number of people involved. This brings around another consequential phenomenon: the publication process of large sequencing studies is extensive and calls for new publication strategies, such as manuscript preprints, which has also been the case for the Genome Aggregation Database. At the time of writing, this holds the data for 125,748 exomes and 15,708 genomes from the same number of individuals (Karczewski*, et al* 2019).

In this review, we focus on a decade of targeted WES in the perspective of malignant hematology. We thematically revisit some of the significant discoveries and niches that use next-generation sequencing (NGS), and we briefly elucidate what WES contributed to and how. As an example, it is now realized that many mutations are shared between the lineages, and some malignancies share both characteristics of the myeloid and lymphoid phenotypes. Moreover, we will touch on how evidence is gathered on the single-cell level in a time when prices of whole genome sequencing push the limits for what can be achieved in research and in the clinic. WES and its use must thus be put into the context of particular research, clinical problems and other sequencing methods to fully appreciate its contribution (Fig 1).

Portraying the exome

Before delving further into the role sequencing has played in elucidating derailed hematopoiesis, we will briefly focus on more technical considerations in a historical context,

which is relevant when evaluating the sequencing studies. Most often, WES is compared to whole genome sequencing (WGS) and targeted panel or amplicon sequencing. It is also compared to chain-terminating dideoxy sequencing, led by Sanger (Sanger*, et al* 1977), which is still in use in clinical laboratories after more than four decades. Only a few years ago, results from next-generation sequencing were expected to be positively confirmed by Sanger sequencing, which often had a poorer sensitivity. Whole exome sequencing is the *targeted* sequencing of *short* reads (Fig 2). These two attributes are important, as they influence quality and possibilities in research and clinically applied analyses – and they represent both its strength and Achilles heel.

The coupling of exome capture arrays to Illumina's sequencing platform in 2009 marked the *hot start* of WES (Choi*, et al* 2009). Armed with a protocol for microarray hybridization followed by sequencing by synthesis on the latest NGS offshoot (Genome Analyzer II, Illumina, San Diego, CA, USA), Ng and colleagues at the University of Washington (Seattle, USA) managed to capture and sufficiently sequence approximately 26 megabases (Mb) of the coding genome, based on twelve individuals (Ng*, et al* 2009). This figure is assessed relative to the approximately 30 Mb size of the exome, corresponding to approximately 1% of the human genome. It was evaluated that the detection of SNPs in the coding genome had equivalent sensitivity to that of WGS (Ng*, et al* 2009). Not long after, the arrays were replaced by in-solution capture of the coding genome (Bainbridge*, et al* 2010), showing reproducibility between technical replicates and a library preparation protocol similar to the ones currently used. The protocol was based on the NimbleGen system (Roche, Basel, Switzerland). While not the first paper describing the hybridization in solution instead of a microarray, a previously performed study only targeted a smaller fraction of the complete exome (Gnirke*, et al* 2009).

In hematology, it soon became clear that even if the control-paired whole-genome resequencing of a cancer was feasible, in technical and economic terms, it was not necessarily practical. Thus, the first implementations of NGS in hematology can be characterized as pseudoexome sequencing. In malignant hematology, the foremost usage of sequencing is to detect somatic mutations, and to do so, the best conditions are (1) evenly distributed reads throughout the regions of interest, (2) as high depth of coverage as possible of both the

malignant sample and the paired control sample, and (3) as low an error rate as possible to avoid false-positive mutations. A random DNA library will, in theory, provide evenly distributed sequencing depth (Lander and Waterman 1988). However, this is a coarse approximation in regard to exome sequencing. Due to the exon target capture being based on hybridization to oligonucleotide probes, the resulting reads distribution for each exon will, at best, be bell-shaped, with the highest number of reads in the center of the exon, leading to differences in statistical power and downstream resolution. The concern was, among other places, raised in PNAS in 2015 under the conclusive title that *"Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants"* (Belkadi*, et al* 2015). It also concluded that WES is not reliable for the detection of copy number variation (CNV). Although such detection poses different challenges than relying on WGS, this is an oversimplification and is influenced by several factors (see (Fromer*, et al* 2012, Hansen*, et al* 2019, Sathirapongsasuti*, et al* 2011) for examples). While we will not go further into details on CNV detection, it must be noted that the Broad Institute recently released a stable pipeline and *best practices* documentation for germline and somatic CNV calling (GATK 4.1, Broad Institute, Cambridge, MA, USA). This pipeline relies on a provided *panel of normals*, such as 40 exome sequenced samples without CNVs to *denoise* the sample of interest.

Collectively, it is now commonly known that exome capturing and PCR introduce bias (Belkadi*, et al* 2015, Braggio*, et al* 2013, Meynert*, et al* 2014, Parla*, et al* 2011, Warr*, et al* 2015). This affects one of the most important features of WES, namely, variant allele frequencies. It is an important point to raise, as sequencing projects often deduce clonal architecture and burden from these data.

Implementations of exome sequencing in hematology

Several themes in hematology can be identified from the last decade, including recurrent mutations, molecular stratification of diagnoses, premalignant lesions and age-related clonal hematopoiesis, allelic burden and detection of residual disease, intrapatient heterogeneity, and demonstration of subclonal mutations. Early signs of implementing sequencing of the coding genome in the field of hematology, and the wealth of information to come, became evident with the first whole cancer genome sequencing of a patient suffering from acute myeloid leukemia (AML) with a normal karyotype (Ley*, et al* 2008). This study preceded the year

when capture-based exome sequencing entered the scene and was followed by another case of cytogenetically normal AML (Mardis*, et al* 2009). Here, the most important finding was that the initially identified somatic mutation in *IDH1*, which substitutes arginine at codon 132, could be observed in 16/188 additional cases by amplicon sequencing. From a technical stance, the correlation of sequenced variant allele frequencies, partly from DNA and reverse-transcribed mRNA, was noteworthy. The first sequencing of an AML patient already suggested subclonal accumulations based on allele frequency, hereby indicating that *FLT3*-ITD was acquired as a late mutation (Ley*, et al* 2008). Many of the concepts used today have existed since very early studies and have been further elaborated and expanded.

The general adoption of NGS in hematology was perhaps not quick to reach clinical laboratory diagnostics, but it was certainly placed in an advantageous position for some of the malignancies compared to other cancer forms. This is partially because of the relative ease of sampling in leukemia and leukemic phase lymphomas, established routine bone marrow aspirations, etc. Furthermore, emergent research involving NGS has rested on already established workflows with regards to storage of DNA and RNA in biobanks, frequent follow-up for remission, relapse evaluation and so on.

Clonal hematopoiesis of healthy and diseased individuals

While hematological malignancies arise from clonal hematopoiesis, one may argue that the line between cancer and noncancerous proliferation has perhaps become even more diffuse with the information gathered from NGS studies. A recent study in the New England Journal of Medicine concluded that mutations persisted in approximately half of the investigated AML cohort (221/430) at complete remission (Jongen-Lavrencic*, et al* 2018). The persistence of clonal mutations may be attributed to residual disease as well as caused by age-related clonal hematopoiesis (Genovese*, et al* 2014, Jaiswal*, et al* 2014). The latter is reinforced by the observation that lingering mutations in *DNMT3A*, *TET2* or *ASXL1* do not confer a higher rate of AML relapse. The current definition of clonal hematopoiesis of indeterminate potential (CHIP) is generally defined as more than 2% variant allele frequency (Steensma*, et al* 2015) of driver genes (see reviews (Sano*, et al* 2018, Steensma 2018)) – a very low threshold for WES.

It is now known that CHIP is a common phenomenon in elderly people, occurring in 10% above 65 years of age without any manifestations of a blood disorder (Genovese*, et al* 2014). To avoid bias from false-positive variants arising from technical errors and germline variants, the putative somatic mutations in the mentioned study were selected from an intermediate allele fraction. The three most frequent and apparently mutated genes were *DNMT3A*, *ASXL1* and *TET2*, which are most frequently seen in myeloid malignancies. Interestingly, Jaiswal *et al* performed low allele frequency screening of 17,182 exomes without paired normal controls (Jaiswal*, et al* 2014) but instead used a panel of controls and a focused cancer variant list from the COSMIC database. From a technical point of view, false-positives are expected to occur in a substantial proportion of the variant calls when approaching a lower variant allele frequency (VAF, threshold defined as 3.5% for single-nucleotide variants (Jaiswal*, et al* 2014)) – an indisputable problem in WES. Back then, the applied somatic caller, MuTect, only processed single-nucleotide variants, so small indels had to be detected by other means. The third study, published the same year, had a similar approach (Xie*, et al* 2014) based on malignant and matched controls from The Cancer Genome Atlas but divided the VAFs into two categories; 2–10% and one above 10%, calling the mutational analysis *unbiased* due to high coverage. However, evidence was gathered to show that WES is not unbiased (Meynert*, et al* 2014, Meynert*, et al* 2013, Parla*, et al* 2011, Warr*, et al* 2015). Although pushing exome sequencing to its extremes is challenging, bioinformatically, the lesson learned in 2014 was clear: clonal hematopoiesis is indisputably linked to the aging genome – whether confounding or directly caused by underlying pathological conditions – and confers an increased risk of developing a hematological malignancy of approximately 1% per year (Genovese*, et al* 2014, Jaiswal*, et al* 2014). The number of included whole exome sequenced samples in these studies was unprecedented in the field of hematology, with the collective effort drawing data from more than thirty thousand screenings.

WGS, WES and targeted panel sequencing have played a large role in elucidating noncancerous clonal hematopoiesis – a phenomenon known long before NGS entered the laboratories and occurring in both lymphoid and myeloid lineages. The continuum of benign clonal expansions to the evolution of a malignant phenotype by accumulating genomic aberrations was readily comprehensible from clinicians' point of view; a direct genomic association between monoclonal B cell lymphocytosis preceding CLL or monoclonal

gammopathy preceding multiple myeloma, etc., became clear. It seems fair to say that no other malignancy in the genomic age has taught hematologists more about cancer heterogeneity, and the longitudinal increment of lesions behind clonal evolution leading a frank malignancy, than CLL. We refer to the recent and thorough review by Crassini *et al* on the molecular pathogenesis of CLL (Crassini*, et al* 2019). This paper also ponders on an important prognosticator where WES falls short, namely immunoglobulin gene rearrangement.

Mapping susceptibility towards clonal lymphoproliferative and myeloproliferative disorders

Although it is impossible to cover all the nooks and crannies of exome sequencing and malignant hematology during the last decade, one research area on the borderline of malignant and benign leukocytosis deserves to be mentioned: the genetic predisposition towards leukemia. The same year as the first publications on exome sequencing entered the stage, a short paper entitled *"Familial Chronic Lymphocytic Leukemia: What Does it Mean to Me?"* aimed to raise awareness of the genetic susceptibility and to guide practitioners (Slager and Kay 2009). Although overall low, familial CLL and the risk of developing lymphoproliferative diseases had been known for many years (Houlston*, et al* 2003), but the genetic etiology was yet unknown. As a concurrent study had just disproved the association between a suspected polymorphic variant in *CXCR4* (Crowther-Swanepoel*, et al* 2009) based on large-scale Sanger sequencing, the problem was posed for genome-wide association studies (GWAS). Nevertheless, the first GWAS utilized arrays and not NGS. Several impressive studies were conducted, which identified risk loci for CLL, such as in *BCL2*. One interesting finding was that several of the variants were common in the population and were thus detectable by SNP arrays (Berndt*, et al* 2013, Crowther-Swanepoel*, et al* 2010, Slager*, et al* 2012). In addition, some of the loci were outside coding regions and, thus, not suited for exome screening. In contrast, the role of mutated *POT1* as a frequent contributor in the development of CLL was initially discovered by WES (Quesada*, et al* 2011, Ramsay*, et al* 2013) and shortly thereafter extended to also hold a susceptibility locus by GWAS (Speedy*, et al* 2014). Because more recent studies focus on somatic variants, the initial genotyping of cases in relation to heritable risk has been extremely important in the *struggle* to understand the complex leukemogenesis of CLL seen from single cases (as exemplified by (Hansen*, et al*

2015)). As with CLL, multiple myeloma is characterized by a preceding monoclonal lymphoproliferative disorder, and for this entity, there are now indications of an inherited risk of developing myeloma through the analyses of germline WES data and possibly disease-causing alleles (Scales*, et al* 2017). Not only have lymphoproliferative disorders been significant in mapping familial cases (Goldin*, et al* 2013), but familial myeloproliferative neoplasms and AML have also received attention in the last decade. Myeloproliferative neoplasms are genomically complex and extend beyond *JAK2*, *MPL* and *CALR* mutations. Because outlining hereditary predispositions becomes a lengthy treatise, we refer to the previous publication in this journal on this specific topic (Rumi and Cazzola 2017). More generally, we refer to Houlston RS *et al* for the extensive work on familial cases and predisposition towards hematological malignancies.

CEBPA is a frequently mutated gene in AML but is also known to be one of the genes implicated in familial cases (Pabst*, et al* 2008, Smith*, et al* 2004). One caveat on this driver pertains to the challenges in targeted sequencing (Yan*, et al* 2016), partly owing to its high guanine-cytosine content, which may be problematic in PCR amplification (Mannelli*, et al* 2017); therefore, the findings often requires confirmation by other modalities until the sequencing quality is adequate, as exemplified by Tawana *et al* (Tawana*, et al* 2015). One important lesson from the past ten years is that not all genes are covered equally well or unbiasedly in WES due to capture, amplification or sequence homology. Examples of the latter are members of the Tet methylcytosine dioxygenase family, e.g., *TET2*, or the most notable members of the Ras GTPases, *NRAS*, *KRAS* and *HRAS*. Other important lessons from WES studies have involved Down syndrome and the predisposition towards both myeloid and lymphoid malignancies (Nikolaev*, et al* 2013, Schwartzman*, et al* 2017, Vesely*, et al* 2017, Yoshida*, et al* 2013).

Sequencing branched into the myeloid lineage

Few cancer forms have received the same amount of attention as acute myeloid leukemia. Not only have recurrent mutations in the initiation of leukemogenesis been elucidated extensively but so have premalignant genomic lesions (Welch*, et al* 2012), which now receive attention in the form of persistent mutations in patients in remission, as demonstrated by targeted sequencing.

Going back to the beginning of next-generation *coding sequencing* in hematology, the focus on coding variants was reasonable, as these variants were already picked as the top tier when evaluating somatic mutations (Mardis*, et al* 2009), even when whole genome sequencing was performed. NGS already had a steady hold in hematology by the time the first exome sequencing studies began to appear. This is perhaps best exemplified by Mardis *et al*, who sequenced almost the complete genome of a patient suffering from AML with minimal maturation and succeeded in showing that the mutations found also occurred recurrently in a cohort of no fewer than 188 patients. One such mutation was in the *IDH1* gene, ubiquitously found in myeloid malignancies, along with previously identified mutations in *NRAS* and *NPM1*. This spurred the hunt for new aberrations by other groups and their mutational frequency in AML. The restriction of downstream analyses to coding regions was also the elaborate strategy used the previous year, in 2008, when the first whole cancer genome was sequenced (Ley*, et al* 2008).

Collectively, sequencing of AML has been an excellent example of a joint scientific effort. The new possibilities arising from sequencing led to bursts of activity in characterizing new genes involved in leukemogenesis across multiple diagnoses, such as *SRSF2* or *SF3B1* (Meggendorfer*, et al* 2012, Yoshida*, et al* 2011). Shortly after proving the involvement of *SF3B1* in myelodysplastic syndromes (MDS) with increased ring sideroblasts (Yoshida*, et al* 2011), it was shown to confer reduced cytopenia and a good prognosis in MDS (Papaemmanuil*, et al* 2011). It was also found to be recurrent in CLL as well but was, in contrast, associated with poor overall survival (Quesada*, et al* 2011).

The blood-forming stem cells in the bone marrow accumulate genomic lesions, leading to senescence or even cancer. As these cells proliferate to replenish the different cell types constituting the blood and immune system, with an estimated frequency of once every 25–50 weeks (Catlin*, et al* 2011), the early mutations propagate along the maturation of the lineages. The fact that the hematopoietic stem cells (HSCs) of the bone marrow display an increasing number of somatic mutations as a result of aging is now indisputable. Equipped with exomes of hematopoietic stem/progenitor cells from individuals in different age groups, Welch and colleagues were able to assess the underlying kinetics with a mutation rate of 0.13 coding

mutations per year (Welch*, et al* 2012) – sensible in the view of the following massive sequencing studies and a practical mental cue for the observed median number of 13 coding mutations in de novo AML (Cancer Genome Atlas Research*, et al* 2013).

The emerging use of exome sequencing could have forecasted the demise of panel sequencing. This, however, was not the case, and indeed, one may argue that recent years have not set the stage for WES but that of targeted sequencing of driver genes in myeloid or lymphoid malignancies. Much work has focused on fortifying the earlier results and elucidating low-frequency variants in residual disease. While the number of driver genes directly involved in the leukemogenesis is relatively small and is thus a comprehensible catalogue for the persistent hematologist, the possible combinations – including positions and types of aberrations – are staggering (Cancer Genome Atlas Research*, et al* 2013, Papaemmanuil*, et al* 2016). It is now known that such signatures identify molecular subgroups. Some mutations are found to be narrowly defined hotspots, such as in *IDH1* or *IDH2*, while others are scattered over a large area, such as *TET2* mutations. The latter observation justifying the wish for comprehensive sequencing as predefined panels will only cover a fraction of the need. After a period hallmarked by productivity but before WES became mainstream, many researchers and clinical laboratories turned to panel sequencing. The seemingly little new knowledge to be gained by exome sequencing and the difficulties in comprehending the broad data output for the clinicians emphasized the usefulness of offering panel sequencing for diagnostic purposes (Roug*, et al* 2014). The focused panels enabled much deeper sequencing, and the overlapping genomic aberrations across disease entities in the lymphoid or myeloid lineages and previous studies made it possible to characterize the genomic profiles and classify patients into molecular or prognostic subgroups, as has been done for MDS (Haferlach*, et al* 2014). In this particular study, investigating the landscape of lesions in 944 patients with a panel of 104 genes, a whole exome approach would have provided poor subclonal resolution and may not have provided much more information for the specific task based on preexisting extensively characterized drivers (Papaemmanuil*, et al* 2013).

Even though NGS was thought to clarify the molecular characterization, making it easier to classify the different diagnoses within malignant hematology, which to some extent it has,

this not turned out to be a straightforward task. Consequently, *heterogeneity* or clonal heterogeneity has been one of the buzzwords of the last decade. Even a readily identifiable form of leukemia, such as acute promyelocytic leukemia (APL), presented itself as a heterogeneous disease (Ibanez*, et al* 2016). Not only has WES differentiated patient-specific sets of mutations with known alterations from other leukemias or lymphomas, but microRNA is also altered in APL (Cancer Genome Atlas Research*, et al* 2013). Therefore, exome sequencing only refers to a small part of the aberrant genome in holistic characterization of the hematological neoplasm. This has been unquestionably notable in the lymphoid diseases, which are marked by a high degree of molecular heterogeneity.

Some major contributions in genomic profiling of the lymphoid malignancies

While the first cancer genome sequenced was from a case of AML and the myeloid neoplasms have received much attention in the last decade, no lymphoma has undergone the same scrutiny as the diffuse large B cell lymphomas (DLBCL, Fig 3). Although several recurrent mutations were known before the advent of exome sequencing, its genomically heterogeneous nature has required large analysis cohorts, such as those in the recent paper published in the New England Journal of Medicine last year, where mutational and expressional profiles were coupled to identify the subgroups and differentiation of DLBCL (Schmitz*, et al* 2018). When considering the expected number of somatic mutations in DLBCL, which is somewhat higher than in AML and other hematological malignancies (Lawrence*, et al* 2014, Lohr*, et al* 2012, Tokheim*, et al* 2016, Zhang*, et al* 2013), it is a suitable candidate for machine learning techniques in computer-aided profiling and diagnostics. In contrast, the powerful approach by Reddy et al, portraying the landscape of DLBCL in a cohort 1,001 patients by WES and RNAseq with functional genomics (Reddy*, et al* 2017), showed just close to 8 mutations per case and 150 putative driver genes, collectively. Interestingly, the paper also concluded that CRISPR-Cas9 knockout of several therapeutically targetable genes, such as *NOTCH2*, did not alter the growth of DLBCL. This may indicate that driver redundancy makes precision medicine a tricky endeavor in some cases. Another quite recent study also showed that the mutational profiles from NGS can be implemented as a predictor for subtype classification (Schmitz*, et al* 2018). The clear advantage of such an approach is that such mutational predictors used in clinical settings may be less sensitive to fluctuations in expression patterns due to sample processing, laboratory

batch effects and computational analyses between laboratories when compared to RNA expression profiles. More rare lymphomas, such as mantle cell lymphoma, show that there is still a need for systematic molecular characterization, where NGS will continue to play a major role.

Multiple myeloma (MM), which is caused by cancerous mature plasma cells, is a very heterogeneous disease. Part of this heterogeneity may be directly attributed to differences in sampling (Sanoja-Flores*, et al* 2018) and cellular impurities. Thus, the malignant cells or plasma cells with monoclonal gammopathy of undetermined significance must be consistently positively selected, e.g., CD138+, from peripheral blood (Lohr*, et al* 2016) or bone marrow. While it is acknowledged that a large part of the driving somatic alterations either occurs in or directly involves the coding regions of the genome, it is also generally accepted that sequencing focusing on these parts can only reveal the molecular oncogenesis to a limited extent. This realization is emphasized by the first large-scale study based on WGS and WES, which identified 35 nonsynonymous somatic point mutations on average, which is in stark contrast to the estimated number of genome-wide mutations of in the thousands (Chapman*, et al* 2011). Multiple myeloma is noted to be the second most common hematological cancer form in adults relative to Non-Hodgkin's lymphoma as a single entity; thus, it is important to elucidate the aberrant molecular pathways of this plasma cell disease. Apart from the conclusion that the underlying mutational patterns of myeloma harboring t(4;14) and t(11;14) are distinct (Walker*, et al* 2012), a previous study also shed light on the clonal diversity and kinetics based on the results of exome sequencing and single-cell genotyping. Importantly, the variant allele frequencies of WES were in striking agreement with the fractions observed from the single cell, where the *ATM* kinase was mutated in all three subclones, probably indicating an early step in oncogenesis. Multiple myeloma displays few highly recurrent genes (Bolli*, et al* 2014), such as *KRAS*, *NRAS* and *BRAF*, in one-third of the patients, and mutations in these genes are considered to be mutually exclusive (Walker*, et al* 2012). Notably, some discrepancies between exome and whole genome data were identified in the aforementioned study.

WES has not only provided a tool for performing large scale screening, but it has also often played an indirect role in the insight into the genomic landmarks of individual hematological

neoplastic entities. One such example is a diagnostically robust marker, the *BRAF* mutation (p.V600E), which is known from other cancer forms such as melanoma. A single case of hairy cell leukemia (HCL) sequenced along with paired mononuclear control cells was enough to initiate the investigation of mutational frequencies in HCL. This was also a time when NGS was validated by Sanger sequencing; thus, the findings were confirmed and fortified by the additional observation in 46/46 other HCL patients (Tiacci*, et al* 2011). Exome sequencing has also had a tremendous impact on the molecular understanding of acute lymphoblastic leukemia (ALL), which is difficult to capture in a brief review. The recurrent mutations of both B- and T-cell ALL bear markers, which otherwise are indicative of a myeloid disorder (Ding*, et al* 2017, Liu*, et al* 2017, Zhang*, et al* 2012), such as *FLT3*, along with classical lymphocytic pathway members. It is suggested that *FLT3* mutations point to an immature pluripotent prothymocyte and that such patients with stem cell-like leukemia may benefit from tyrosine kinase inhibitors (Neumann*, et al* 2013a), which may improve their prognosis (Zhang*, et al* 2012).

The degree of complexity has definitely not been reduced with the screening of genomes and exomes. However, while the myeloid and lymphoid branches have been clearly divided in the classical model of hematopoiesis and permeated the clinical view, it is now recognized that certain lesions in immature cells with some degree of pluripotency are capable of giving rise to mixed lineages, biphenotypic leukemias or malignancies stuck in early precursor differentiation. Early T-cell precursor acute lymphoblastic leukemia (ETP-ALL) is a molecular manifestation of such a potential. If one gene were to capture the ambiguous nature of the neoplasms derived from B- and T-lineage lymphoid precursors, the Ikaros transcription factor *IKZF1* is a strong candidate for such a gene, as it is a *master regulator of lymphocyte differentiation*. Although primarily recognized for its role in B-ALL (Marke*, et al* 2018), growing evidence supports that *IKZF1* plays a cardinal role in ETP-ALL. (Hansen*, et al* 2018, Zhang*, et al* 2012) Additionally, germline variants have been investigated in the predisposition of developing childhood acute lymphoblastic leukemia (Churchman*, et al* 2018). Other mutational markers of ETP-ALL have clear-cut overlaps with the myeloid leukemias, such as *DNMT3A (Neumann, et al 2013b)*, *FLT3* or mutations in the JAK-STAT pathway.

One latecomer in sequencing is classical Hodgkin's lymphoma (cHL) (Mata*, et al* 2017, Rotunno*, et al* 2016). The molecular pathogenesis is still not fully charted, which may partially have been influenced by improved prognosis. This may have also been affected by the historical ease of diagnosis from the view of a hematopathologist: those which are Hodgkin's, and the others *which are not* – a classification different from that used in genomics. One specific challenge in genomic characterization of cHL is the small fraction of malignant cells in the bulk tumor mass. Because WES is not directly geared for detection of 1% malignant cells, the methodology calls for cell sorting. By using an optimized exome library preparation Reichel *et al* could perform sequencing on 10 nanograms of Hodgkin and Reed-Sternberg cell DNA and thereby obviate the need for whole genome amplification (Reichel*, et al* 2015).

Profiling rare subclones and low-frequency malignant cells down to the single-cell level

One of the advances of the last decade has been single-cell analysis, which is – at least in theory – well-suited for the characterization of stem cells. It was realized that using WES for the identification of mutational fingerprints was an excellent base for the development of techniques with a high enough resolution for the evaluation of low burden malignancy, such as the quantification of minimal or measurable residual diseases in patients with clinical remission. While WES is suited for the identification of mutation in driver genes, once identified, focused deep sequencing in combination with cell sorting may be implemented to pinpoint rare malignant cells (Malmberg*, et al* 2017). Despite major efforts, few concepts in hematology are as pervasive and yet so vaguely defined as the hematopoietic stem cell. However, the very concept of stemness is extended to the self-renewal and differentiation capacities of the leukemias (Jordan 2007). Thus, the single-cell analyses fit the proposal that the transition of the HSC to frank leukemia is defined by preleukemic lesions of the stem cell. An accumulation of somatic mutations and clonal progression of de novo AML were analyzed even before the genomic landscape of such cases was mapped (Jan*, et al* 2012). Exome sequencing became highly relevant in the search for single leukemic cells, as the genomes of these malignant cells, which, in contrast to normal cells, are capable of self-renewal, are marked by lesions that are often in the coding regions of the genome (Jan*, et al* 2012). The subject is only briefly touched upon, as these efforts more often belong to transcriptome sequencing (Lai*, et al* 2018, Yashiro*, et al* 2009), which has also been utilized

in profiling clonal architecture in childhood ALL by single-cell sequencing (SCS) (Gawad*, et al* 2014).

Foremost, the sequencing studies have been used to delineate the occurrence, prevalence and successions of genomic lesions in hematological malignancies in the context of myeloid and lymphoid lineages. Together with the maturation of single-cell analyses, it has, in principle, become possible to characterize the shared hematopoietic stem cell (Baron*, et al* 2018) by parallel sequencing, thereby coming closer to portraying the immature cells and early progenitors. Nevertheless, researchers in the field of hematology have been puzzled by the many occurrences of concurrent or successive myeloid and lymphoid malignancies, which otherwise have been marked by a wide cleft between the lineages in earlier research. To a large extent, this has been driven by sequencing technologies developed in the last two decades, whereas the usage of targeted exome sequencing is confined to the last ten years.

Next-generation sequencing has helped define and shape the field of single-cell research, and some of the most important publications using SCS have focused on specific problems in hematology. As early as 2012, a study identified mutations in both leukemic cells and preleukemic hematopoietic stem cells based on the results from an initial exome screening of sorted cells and paired CD3+ control (Jan*, et al* 2012). Instead of whole genome amplification, which is largely the current method in SCS, the cells were sorted into plates with subsequent culturing and genotyping. By the time the first report on whole genome sequencing was published, another paper on advances and patents in NGS briefly stressed the importance of developing single-cell sequencing technologies and described the potential amplification of a whole genome by multiple displacement amplification (Lin*, et al* 2008). The low signal-to-noise and allelic drop-outs still pose a challenge in SCS studies (Alves and Posada 2018, Gawad*, et al* 2016, Simonsen*, et al* 2018), and capture-based targeted sequencing may suffer even more than whole genome or RNA sequencing due to its additional steps in sample processing. As single-cell sequencing evidently has several issues to tackle, such as biased capture, partial or complete drop out of targeted regions, and base substitutions, transcriptome sequencing has become a popular alternative. Whereas transfer of information from the two copies of DNA in a single cell is a delicate process, sequencing a wealth of transcripts is statistically attractive.

Whereas a large number of patients were included in some of the first WES studies, the current studies may also focus on a few individuals with a large throughput of single cells with state-of-the-art microfluidics (Pellegrino*, et al* 2018) or sorting with subsequent whole-genome amplification (Walter*, et al* 2018).

Discussion

Next-generation sequencing has helped cancer researchers to understand how clones evolve and that subclones often exist side by side. Solid tumors are found at the most extreme end, with intratumoral heterogeneity reflected by different sets of spatially confined somatic mutations (Gerlinger*, et al* 2012). In the timespan between the 2008 and the 2016 World Health Organization (WHO) *classification of hematopoietic and lymphoid tumors,* a riveting transition took place: the field of hematology entered the genomic era.

Whole exome sequencing, as a technology, has contributed immensely to the field of hematology in the last ten years, and many of the pivotal findings were presented within the first five years of its existence. These concepts matured in the following years, and themes changed to highly specific niches, such as single-cell sequencing, the field of clonal hematopoiesis of indeterminate potential and detection of minimal residual disease, to mention a few. However, it is highly plausible that WGS will gradually, and perhaps completely, replace WES during the next decade. This is already occurring, and there are several factors influencing this development. First, it may be argued that exome sequencing is not inexpensive. In addition to proprietary kits, the protocols require additional manual labor. Although the costs involved in capture-based exome sequencing have also decreased in recent years, WGS has gained the largest advantage. In 2008, the details of the genome of James D. Watson, co-discoverer of the DNA double helix structure, were published (Wheeler*, et al* 2008). Sequenced the previous year and completed in two months, the estimated price was less than $1 million – a price tag that has since decreased a hundred- to a thousand-fold. Effectively, WGS is catching up to its smaller cousin, now that the $1,000 genome with 30x coverage is readily available. Second, technical bias is introduced through exon capture and target amplification by polymerase chain reaction (PCR), which is also an issue with panel sequencing, but not in WGS protocols. Simply increasing the sequencing depth for a sample,

which is highly feasible with the current low cost per sequenced base, does not alone alleviate the problems of false-positive variant detection. As sequencing offers a tool for the detection of low-frequency subclones or measurable residual disease detection and characterization in otherwise complete remission, reducing technical bias, sample cross-talk or noise is highly relevant for its clinical applicability. Adding unique molecular identifiers (UMIs) to the DNA library will help to determine and map the extent of PCR bias in a sequenced sample (Best*, et al* 2015, MacConaill*, et al* 2018), but this extension is still not commonly adopted in library preparation or downstream bioinformatics. As UMIs have been used and studied extensively in RNA and single-cell sequencing (Sena*, et al* 2018), this will likely become mainstream in cancer diagnostics for increased precision in targeted and whole genome sequencing. The general consensus that whole genome sequencing is superior to exome sequencing in terms of quality is currently forming (Wang*, et al* 2017), but WGS still suffers from a shallow sequencing depth, which is typically one-third that of WES or even lower. In comparison, a site-specific depth of panel or amplicon sequencing may be two to three orders of magnitude higher, and consequently, observations of a few hundred are often discarded to minimize false-positive variant calls arising from PCR, among other considerations.

While depth of coverage is used as a quality measure of WES it does not capture the variability of sequencing reads along the exome, which may be detrimental for clinical testing. Reporting that a certain proportion of the bases, e.g. 90%, reach the specified threshold is useless if relevant oncogenes, such as *CEBPA*, only attain sporadic coverage at critical sites owing to a high GC percentage (Kong*, et al* 2018, Roy*, et al* 2018), poor primer design etc. Currently, several library preparation kits exist, such as SureSelect Clinical Research Exome (Agilent, Santa Clara, CA, USA) or SeqCap EZ MedExome (Roche, Basel, Switzerland), which combines the genomic breadth of WES with a higher depth of coverage in medical relevant regions. These extensions to classical targeted WES may thus provide higher resolution in somatic variant calling, copy gains, deletions, and other structural variants when needed.

An argument against the adoption of WGS is the quantity of data being generated and its bioinformatics challenges. In addition, in many cases, the tumor DNA and the germline DNA of the patient are both sequenced for a better bioinformatics workup. However, this strategy

may change due to improved pipelines and databases for unpaired samples. One suggestion to circumvent such bottlenecks of data handling is the WGS postsequencing retrieval of coding regions for initial analysis. As the performance of a vendor's exome kit can only be guaranteed in targeted regions, some exclusion of off-target reads is already performed by the sequencing centers or bioinformaticians. In some aspects, RNA transcriptome sequencing also offers a direct alternative to WES, as *i)* it covers the coding regions, *ii)* it is feasible for the detection of somatic mutations, *iii)* RNA sequencing provides additional information from the transcriptional profile of the malignancy, and *iv)* the workflow is simple and requires a small amount of input material. However, the additional information provided by DNA sequencing, such as CNVs and allele frequency distributions, is partly or completely lost.

Short-read sequencing has been a huge market, but this may change, as several opportunities arise from long-read sequencing, such as identifying chromosomal translocations and chromosomal phasing of variants. In addition, short reads give rise to problems with sequencing of repetitive stretches of DNA and structural variants, and generally fail in detecting breakpoints of chromosomal translocations. In the next decade, we will probably experience a revival of long-read sequencing in the form of 3rd generation sequencing such as that provided by Oxford Nanopore Technologies (Oxford, UK).

Concluding remarks

The replacement of WES by sequencing of the full genome seems inevitable. This "return" to whole genome sequencing represents a full circle, with sequencing of the first cancer genome as the starting point of the exome sequencing revolution, and its early signs of senescence a decade after the first WES studies. The adept usage of WES and WGS to chart oncogenic drivers led to the adoption of figurative jargon, describing the landscapes of malignancies and capturing researchers' perception of new possibilities and its returns. Undoubtedly, WGS will steadily be implemented in routine diagnostics in the next decade.

Author contributions

MCH drafted the paper and figures based on a shared initiative, and all authors contributed substantially to the final manuscript.

Disclosures

MCH is part-time researcher and independent consultant in NGS and bioinformatics. TH is part-owner of MLL Munich Leukemia Laboratory. CGN has nothing to disclose.

Acknowledgements

The authors would like to thank Dina Mohyeldeen, MD, and Vickie S. Kristensen for their critical view of the final draft.

References

Alves, J.M. & Posada, D. (2018) Sensitivity to sequencing depth in single-cell cancer genomics. *Genome Med,* **10,** 29.

Bainbridge, M.N., Wang, M., Burgess, D.L., Kovar, C., Rodesch, M.J., et al. (2010) Whole exome capture in solution with 3 Gbp of data. *Genome Biol,* **11,** R62.

Baron, C.S., Kester, L., Klaus, A., Boisset, J.C., Thambyrajah, R., et al. (2018) Single-cell transcriptomics reveal the dynamic of haematopoietic stem cell production in the aorta. *Nat Commun,* **9,** 2517.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., et al. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A,* **112,** 5473-5478.

Berndt, S.I. & Skibola, C.F. & Joseph, V. & Camp, N.J. & Nieters, A., et al. (2013) Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nat Genet,* **45,** 868-876.

Best, K., Oakes, T., Heather, J.M., Shawe-Taylor, J. & Chain, B. (2015) Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Sci Rep,* **5,** 14629.

Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., et al. (2014) Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun,* **5,** 2997.

Braggio, E., Egan, J.B., Fonseca, R. & Stewart, A.K. (2013) Lessons from next-generation sequencing analysis in hematological malignancies. *Blood Cancer J,* **3,** e127.

Cancer Genome Atlas Research, N. & Ley, T.J. & Miller, C. & Ding, L. & Raphael, B.J., et al. (2013) Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med,* **368,** 2059-2074.

Catlin, S.N., Busque, L., Gale, R.E., Guttorp, P. & Abkowitz, J.L. (2011) The replication rate of human hematopoietic stem cells in vivo. *Blood,* **117,** 4460-4466.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., et al. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature,* **471,** 467-472.

Choi, M., Scholl, U.I., Ji, W., Liu, T., Tikhonova, I.R., et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A,* **106,** 19096-19101.

Chong, J.X., Buckingham, K.J., Jhangiani, S.N., Boehm, C., Sobreira, N., et al. (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet,* **97,** 199-215.

Churchman, M.L., Qian, M., Te Kronnie, G., Zhang, R., Yang, W., et al. (2018) Germline Genetic IKZF1 Variation and Predisposition to Childhood Acute Lymphoblastic Leukemia. *Cancer Cell,* **33,** 937-948 e938.

Crassini, K., Stevenson, W.S., Mulligan, S.P. & Best, O.G. (2019) Molecular pathogenesis of chronic lymphocytic leukaemia. *Br J Haematol,* **186,** 668-684.

Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M.C., Dobbins, S.E., Torres, M., et al. (2010) Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet,* **42,** 132-136.

Crowther-Swanepoel, D., Qureshi, M., Dyer, M.J., Matutes, E., Dearden, C., et al. (2009) Genetic variation in CXCR4 and risk of chronic lymphocytic leukemia. *Blood,* **114,** 4843-4846.

Ding, L.W., Sun, Q.Y., Tan, K.T., Chien, W., Mayakonda, A., et al. (2017) Mutational Landscape of Pediatric Acute Lymphoblastic Leukemia. *Cancer Res,* **77,** 390-400.

Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., et al. (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet,* **91,** 597-607.

Gawad, C., Koh, W. & Quake, S.R. (2014) Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci U S A,* **111,** 17947-17952.

Gawad, C., Koh, W. & Quake, S.R. (2016) Single-cell genome sequencing: current state of the science. *Nat Rev Genet,* **17,** 175-188.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., et al. (2015) A global reference for human genetic variation. *Nature,* **526,** 68-74.

Genovese, G., Kahler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., et al. (2014) Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med,* **371,** 2477-2487.

Gerlinger, M., Rowan, A.J., Horswell, S., Math, M., Larkin, J., et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med,* **366,** 883-892.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol,* **27,** 182-189.

Goldin, L.R., McMaster, M.L. & Caporaso, N.E. (2013) Precursors to lymphoproliferative malignancies. *Cancer Epidemiol Biomarkers Prev,* **22,** 533-539.

Goryca, K., Kulecka, M., Paziewska, A., Dabrowska, M., Grzelak, M., et al. (2018) Exome scale map of genetic alterations promoting metastasis in colorectal cancer. *BMC Genet,* **19,** 85.

Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., et al. (2014) Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia,* **28,** 241-247.

Hansen, M.C., Cédile, O., Ludvigsen, M., Kjeldsen, E., Møller, P.L., et al. (2019) Integrating detection of copy neutral chromosomal losses in a clinical setting in leukemia and lymphoma by means of allelic imbalance and read depth ratio comparison. *bioRxiv*.

Hansen, M.C., Nederby, L., Kjeldsen, E., Petersen, M.A., Ommen, H.B., et al. (2018) Case report: Exome sequencing identifies T-ALL with myeloid features as a IKZF1-struck early precursor T-cell malignancy. *Leuk Res Rep,* **9,** 1-4.

Hansen, M.C., Nyvold, C.G., Roug, A.S., Kjeldsen, E., Villesen, P., et al. (2015) Nature and nurture: a case of transcending haematological pre-malignancies in a pair of monozygotic twins adding possible clues on the pathogenesis of B-cell proliferations. *Br J Haematol,* **169,** 391-400.

Houlston, R.S., Sellick, G., Yuille, M., Matutes, E. & Catovsky, D. (2003) Causation of chronic lymphocytic leukemia--insights from familial disease. *Leuk Res,* **27,** 871-876.

Ibanez, M., Carbonell-Caballero, J., Garcia-Alonso, L., Such, E., Jimenez-Almazan, J., et al. (2016) The Mutational Landscape of Acute Promyelocytic Leukemia Reveals an Interacting Network of Co-Occurrences and Recurrent Mutations. *PLoS One,* **11,** e0148346.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., et al. (2014) Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med,* **371,** 2488-2498.

Jan, M., Snyder, T.M., Corces-Zimmerman, M.R., Vyas, P., Weissman, I.L., et al. (2012) Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci Transl Med,* **4,** 149ra118.

Jongen-Lavrencic, M., Grob, T., Hanekamp, D., Kavelaars, F.G., Al Hinai, A., et al. (2018) Molecular Minimal Residual Disease in Acute Myeloid Leukemia. *N Engl J Med,* **378,** 1189-1199.

Jordan, C.T. (2007) The leukemic stem cell. *Best Pract Res Clin Haematol,* **20,** 13-18.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., et al. (2019).

Kong, S.W., Lee, I.H., Liu, X., Hirschhorn, J.N. & Mandl, K.D. (2018) Measuring coverage and accuracy of whole-exome sequencing in clinical context. *Genet Med,* **20,** 1617-1626.

Lai, S., Huang, W., Xu, Y., Jiang, M., Chen, H., et al. (2018) Comparative transcriptomic analysis of hematopoietic system between human and mouse by Microwell-seq. *Cell Discov,* **4,** 34.

Lander, E.S. & Linton, L.M. & Birren, B. & Nusbaum, C. & Zody, M.C., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409,** 860-921.

Lander, E.S. & Waterman, M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* **2,** 231-239.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., et al. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature,* **505,** 495-501.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* **536,** 285-291.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature,* **456,** 66-72.

Lin, B., Wang, J. & Cheng, Y. (2008) Recent Patents and Advances in the Next-Generation Sequencing Technologies. *Recent Pat Biomed Eng,* **2008,** 60-67.

Liu, Y., Easton, J., Shao, Y., Maciaszek, J., Wang, Z., et al. (2017) The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat Genet,* **49,** 1211-1218.

Lohr, J.G., Kim, S., Gould, J., Knoechel, B., Drier, Y., et al. (2016) Genetic interrogation of circulating multiple myeloma cells at single-cell resolution. *Sci Transl Med,* **8,** 363ra147.

Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., et al. (2012) Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc Natl Acad Sci U S A,* **109,** 3879-3884.

MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., et al. (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics,* **19,** 30.

Malmberg, E.B., Stahlman, S., Rehammar, A., Samuelsson, T., Alm, S.J., et al. (2017) Patient-tailored analysis of minimal residual disease in acute myeloid leukemia using next-generation sequencing. *Eur J Haematol,* **98,** 26-37.

Mannelli, F., Ponziani, V., Bencini, S., Bonetti, M.I., Benelli, M., et al. (2017) CEBPA-double-mutated acute myeloid leukemia displays a unique phenotypic profile: a reliable screening method and insight into biological features. *Haematologica,* **102,** 529-540.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., et al. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med,* **361,** 1058-1066.

Marke, R., van Leeuwen, F.N. & Scheijen, B. (2018) The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. *Haematologica,* **103,** 565-574.

Mata, E., Diaz-Lopez, A., Martin-Moreno, A.M., Sanchez-Beato, M., Varela, I., et al. (2017) Analysis of the mutational landscape of classic Hodgkin lymphoma identifies disease heterogeneity and potential therapeutic targets. *Oncotarget,* **8,** 111386-111395.

Meggendorfer, M., Roller, A., Haferlach, T., Eder, C., Dicker, F., et al. (2012) SRSF2 mutations in 275 cases with chronic myelomonocytic leukemia (CMML). *Blood,* **120,** 3080-3088.

Meynert, A.M., Ansari, M., FitzPatrick, D.R. & Taylor, M.S. (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. *BMC Bioinformatics,* **15,** 247.

Meynert, A.M., Bicknell, L.S., Hurles, M.E., Jackson, A.P. & Taylor, M.S. (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinformatics,* **14,** 195.

Neumann, M., Coskun, E., Fransecky, L., Mochmann, L.H., Bartram, I., et al. (2013a) FLT3 mutations in early T-cell precursor ALL characterize a stem cell like leukemia and imply the clinical use of tyrosine kinase inhibitors. *PLoS One,* **8,** e53190.

Neumann, M., Heesch, S., Schlee, C., Schwartz, S., Gokbuget, N., et al. (2013b) Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. *Blood,* **121,** 4749-4752.

Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature,* **461,** 272-276.

Nikolaev, S.I., Santoni, F., Vannier, A., Falconnet, E., Giarin, E., et al. (2013) Exome sequencing identifies putative drivers of progression of transient myeloproliferative disorder to AMKL in infants with Down syndrome. *Blood,* **122,** 554-561.

Pabst, T., Eyholzer, M., Haefliger, S., Schardt, J. & Mueller, B.U. (2008) Somatic CEBPA mutations are a frequent second event in families with germline CEBPA mutations and familial acute myeloid leukemia. *J Clin Oncol,* **26,** 5088-5093.

Papaemmanuil, E., Cazzola, M., Boultwood, J., Malcovati, L., Vyas, P., et al. (2011) Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med,* **365,** 1384-1395.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., et al. (2016) Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med,* **374,** 2209-2221.

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., et al. (2013) Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood,* **122,** 3616-3627; quiz 3699.

Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M., et al. (2011) A comparative analysis of exome capture. *Genome Biol,* **12,** R97.

Pellegrino, M., Sciambi, A., Treusch, S., Durruthy-Durruthy, R., Gokhale, K., et al. (2018) High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Res,* **28,** 1345-1352.

Quesada, V., Conde, L., Villamor, N., Ordonez, G.R., Jares, P., et al. (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat Genet,* **44,** 47-52.

Ramsay, A.J., Quesada, V., Foronda, M., Conde, L., Martinez-Trillos, A., et al. (2013) POT1 mutations cause telomere dysfunction in chronic lymphocytic leukemia. *Nat Genet,* **45,** 526-530.

Reddy, A., Zhang, J., Davis, N.S., Moffitt, A.B., Love, C.L., et al. (2017) Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell,* **171,** 481-494 e415.

Reichel, J., Chadburn, A., Rubinstein, P.G., Giulino-Roth, L., Tam, W., et al. (2015) Flow sorting and exome sequencing reveal the oncogenome of primary Hodgkin and Reed-Sternberg cells. *Blood,* **125,** 1061-1072.

Rotunno, M., McMaster, M.L., Boland, J., Bass, S., Zhang, X., et al. (2016) Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene. *Haematologica,* **101,** 853-860.

Roug, A.S., Hansen, M.C., Nederby, L. & Hokland, P. (2014) Diagnosing and following adult patients with acute myeloid leukaemia in the genomic age. *Br J Haematol,* **167,** 162-176.

Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., et al. (2018) Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J Mol Diagn,* **20,** 4-27.

Rumi, E. & Cazzola, M. (2017) Advances in understanding the pathogenesis of familial myeloproliferative neoplasms. *Br J Haematol,* **178,** 689-698.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* **74,** 5463-5467.

Sano, S., Wang, Y. & Walsh, K. (2018) Clonal Hematopoiesis and Its Impact on Cardiovascular Disease. *Circ J,* **83,** 2-11.

Sanoja-Flores, L., Flores-Montero, J., Garces, J.J., Paiva, B., Puig, N., et al. (2018) Next generation flow for minimally-invasive blood characterization of MGUS and multiple myeloma at diagnosis based on circulating tumor plasma cells (CTPC). *Blood Cancer J,* **8,** 117.

Sathirapongsasuti, J.F., Lee, H., Horst, B.A., Brunner, G., Cochran, A.J., et al. (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics,* **27,** 2648-2654.

Scales, M., Chubb, D., Dobbins, S.E., Johnson, D.C., Li, N., et al. (2017) Search for rare protein altering variants influencing susceptibility to multiple myeloma. *Oncotarget,* **8,** 36203-36210.

Schmitz, R., Wright, G.W., Huang, D.W., Johnson, C.A., Phelan, J.D., et al. (2018) Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N Engl J Med,* **378,** 1396-1407.

Schwartzman, O., Savino, A.M., Gombert, M., Palmi, C., Cario, G., et al. (2017) Suppressors and activators of JAK-STAT signaling at diagnosis and relapse of acute lymphoblastic leukemia in Down syndrome. *Proc Natl Acad Sci U S A,* **114,** E4030-E4039.

Sena, J.A., Galotto, G., Devitt, N.P., Connick, M.C., Jacobi, J.L., et al. (2018) Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci Rep,* **8,** 13121.

Simonsen, A.T., Hansen, M.C., Kjeldsen, E., Moller, P.L., Hindkjaer, J.J., et al. (2018) Systematic evaluation of signal-to-noise ratio in variant detection from single cell genome multiple displacement amplification and exome sequencing. *BMC Genomics,* **19,** 681.

Slager, S.L. & Kay, N.E. (2009) Familial chronic lymphocytic leukemia: what does it mean to me? *Clin Lymphoma Myeloma,* **9 Suppl 3,** S194-197.

Slager, S.L., Skibola, C.F., Di Bernardo, M.C., Conde, L., Broderick, P., et al. (2012) Common variation at 6p21.31 (BAK1) influences the risk of chronic lymphocytic leukemia. *Blood,* **120,** 843-846.

Smith, M.L., Cavenagh, J.D., Lister, T.A. & Fitzgibbon, J. (2004) Mutation of CEBPA in familial acute myeloid leukemia. *N Engl J Med,* **351,** 2403-2407.

Speedy, H.E., Di Bernardo, M.C., Sava, G.P., Dyer, M.J., Holroyd, A., et al. (2014) A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nat Genet,* **46,** 56-60.

Steensma, D.P. (2018) Clinical consequences of clonal hematopoiesis of indeterminate potential. *Blood Adv,* **2,** 3404-3410.

Steensma, D.P., Bejar, R., Jaiswal, S., Lindsley, R.C., Sekeres, M.A., et al. (2015) Clonal hematopoiesis of indeterminate potential and its distinction from myelodysplastic syndromes. *Blood,* **126,** 9-16.

Tawana, K., Wang, J., Renneville, A., Bodor, C., Hills, R., et al. (2015) Disease evolution and outcomes in familial AML with germline CEBPA mutations. *Blood,* **126,** 1214-1223.

Tiacci, E., Trifonov, V., Schiavoni, G., Holmes, A., Kern, W., et al. (2011) BRAF mutations in hairy-cell leukemia. *N Engl J Med,* **364,** 2305-2315.

Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. & Karchin, R. (2016) Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A,* **113,** 14330-14335.

Venter, J.C. & Adams, M.D. & Myers, E.W. & Li, P.W. & Mural, R.J., et al. (2001) The sequence of the human genome. *Science,* **291,** 1304-1351.

Vesely, C., Frech, C., Eckert, C., Cario, G., Mecklenbrauker, A., et al. (2017) Genomic and transcriptional landscape of P2RY8-CRLF2-positive childhood acute lymphoblastic leukemia. *Leukemia,* **31,** 1491-1501.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., et al. (2013) Cancer genome landscapes. *Science,* **339,** 1546-1558.

Walker, B.A., Wardell, C.P., Melchor, L., Hulkki, S., Potter, N.E., et al. (2012) Intraclonal heterogeneity and distinct molecular mechanisms characterize the development of t(4;14) and t(11;14) myeloma. *Blood,* **120,** 1077-1086.

Walter, C., Pozzorini, C., Reinhardt, K., Geffers, R., Xu, Z., et al. (2018) Single-cell whole exome and targeted sequencing in NPM1/FLT3 positive pediatric acute myeloid leukemia. *Pediatr Blood Cancer,* **65**.

Wang, Q., Shashikant, C.S., Jensen, M., Altman, N.S. & Girirajan, S. (2017) Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep,* **7,** 885.

Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N., et al. (2015) Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda),* **5,** 1543-1550.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell,* **150,** 264-278.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature,* **452,** 872-876.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., et al. (2007) The genomic landscapes of human breast and colorectal cancers. *Science,* **318,** 1108-1113.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., et al. (2014) Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med,* **20,** 1472-1478.

Yan, B., Hu, Y., Ng, C., Ban, K.H., Tan, T.W., et al. (2016) Coverage analysis in a targeted amplicon-based next-generation sequencing panel for myeloid neoplasms. *J Clin Pathol,* **69,** 801-804.

Yashiro, Y., Bannai, H., Minowa, T., Yabiku, T., Miyano, S., et al. (2009) Transcriptional profiling of hematopoietic stem cells by high-throughput sequencing. *Int J Hematol,* **89,** 24-33.

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., et al. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature,* **478,** 64-69.

Yoshida, K., Toki, T., Okuno, Y., Kanezaki, R., Shiraishi, Y., et al. (2013) The landscape of somatic mutations in Down syndrome-related myeloid disorders. *Nat Genet,* **45,** 1293-1299.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., et al. (2012) The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature,* **481,** 157-163.

Zhang, J., Grubor, V., Love, C.L., Banerjee, A., Richards, K.L., et al. (2013) Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc Natl Acad Sci U S A,* **110,** 1398-1403.

Figures and legends

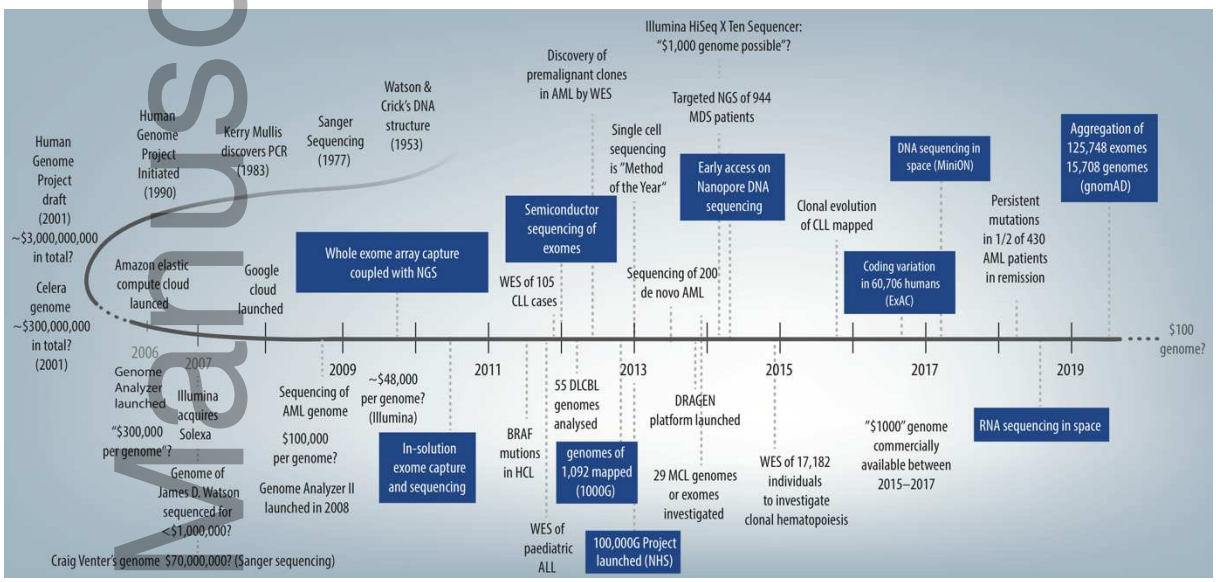


Fig 1) Timeline of postmillennial sequencing with selected milestones and curiosities. Whole exome sequencing only represents a fraction of the developments and achievements in the genomic era. Notably, drafts of the human genome were published in 2001 accompanied by staggering costs. The Genome Analyzer II (Illumina) entered the market in 2008, the same year the Genome Analyzer was used to sequence the first cancer genome from a patient with AML and a year ahead of targeted whole exome sequencing. The price of genome sequencing has decreased 100- to 1000-fold in the last decade, and we will likely experience the first commercially available $100 genomes in the next decade. *Many innovations have been left out due to space limitations.*

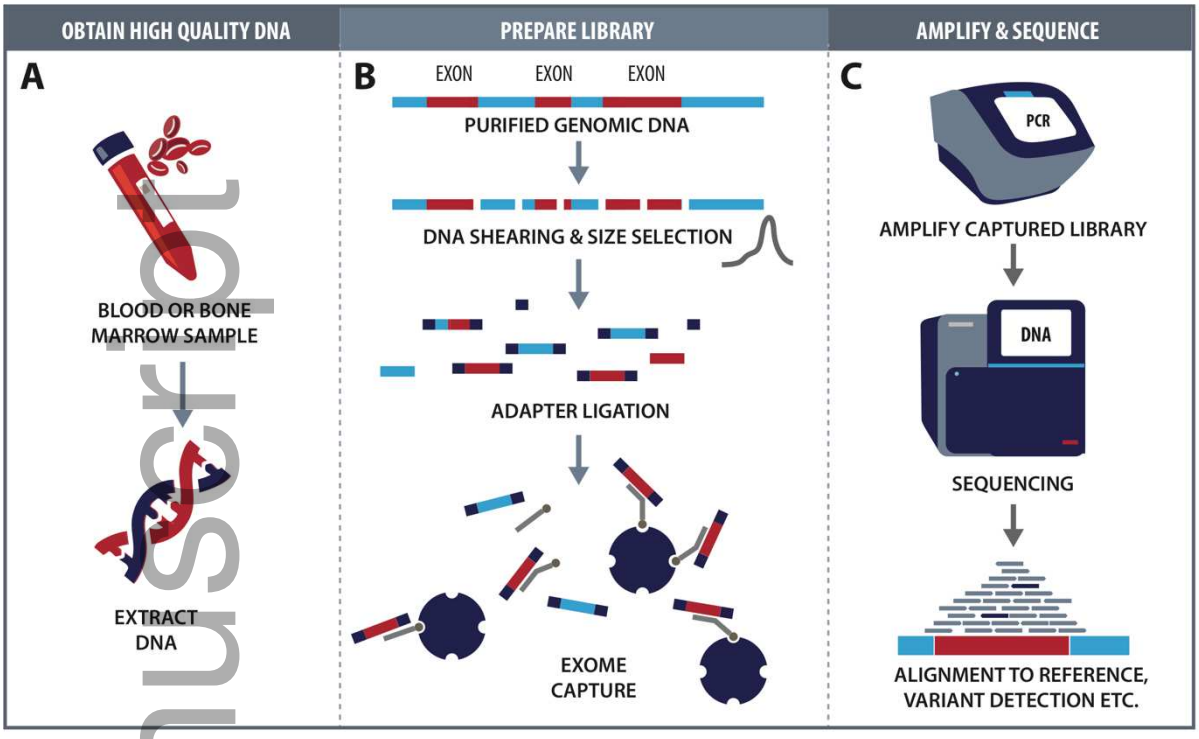


Fig 2) Generalized workflow of whole exome sequencing. The first step is obtaining DNA of sufficient quality (A), as this affects all downstream analyses. Importantly, the results generated from formalin-fixed paraffin-embedded tissue, bulk DNA, sorted cells, etc., are not necessarily compatible with the aim. Library preparation (B) is also a crucial step, where different workflows of shearing, size selection, adapter ligation and capture will affect the final result. PCR amplification is used to enrich targeted DNA before sequencing (C). The researcher must be aware of the potential bias introduced by amplification and be able to handle this. Furthermore, they must know that no two downstream algorithms will output the exact same sets of variants or even copy number variations.

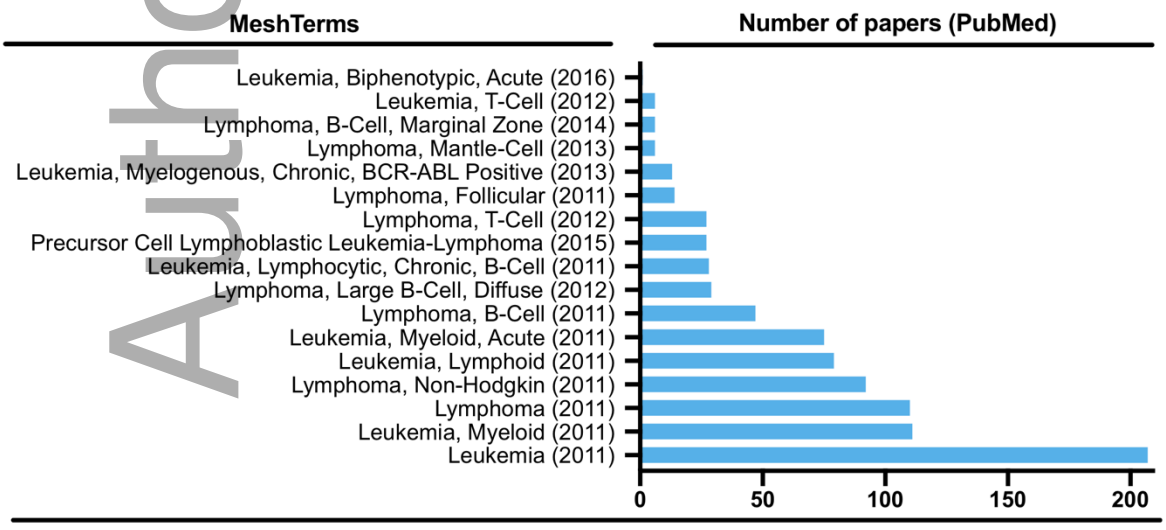
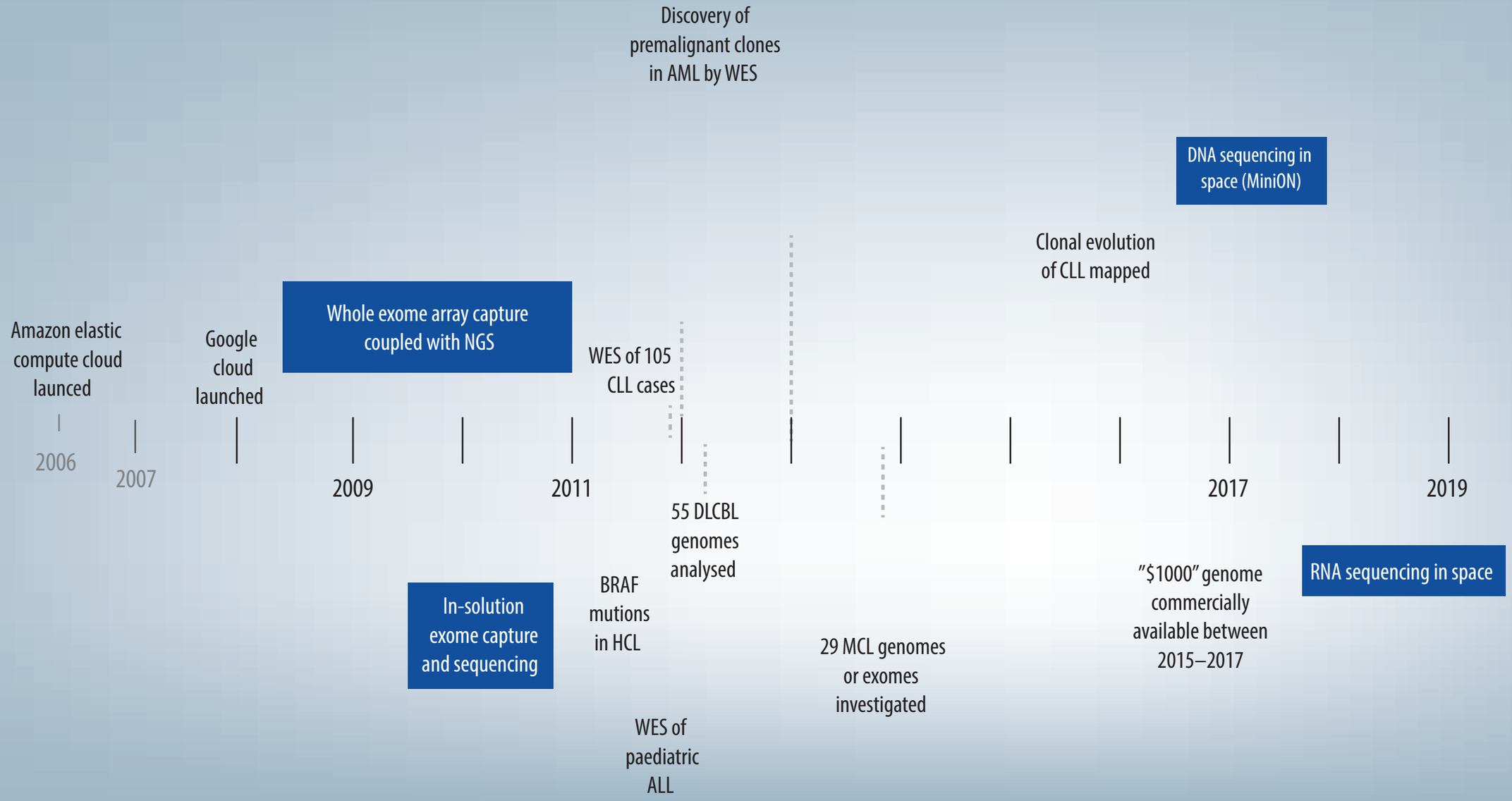


Fig 3) The number of published papers in the PubMed database covering the *selected* MeSH Term subjects and containing the term *exome sequencing* (May 2019). The graph shows the relative abundance and thus focuses on a given hematological malignancy. The first occurrence in the database is shown in parentheses. For comparison, capture-based exome sequencing was published in 2009, and the first cancer genome paper was published in 2008.


Discovery of premalignant clones in AML by WES
DNA sequencing in space (MiniON)
Clonal evolution of CLL mapped
Amazon elastic compute cloud launced
Google cloud launched
Whole exome array capture coupled with NGS
WES of 105 CLL cases
2006
2007
2009
2011
2017
2019
55 DLCBL genomes analysed
In-solution exome capture and sequencing
BRAF mutions in HCL
29 MCL genomes or exomes investigated
"$1000" genome commercially available between 2015–2017
RNA sequencing in space
WES of paediatric ALL