RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL RESEARCH

Current status of use of high throughput nucleotide sequencing in rheumatology

Sebastian Boegel ,[1] John C Castle,[2] Andreas Schwarting[1,3,4]

To cite: Boegel S, Castle JC, Schwarting A. Current status of use of high throughput nucleotide sequencing in rheumatology. *RMD Open* 2021;**7**:e001324. doi:10.1136/rmdopen-2020-001324

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2020-001324).

Received 13 May 2020
Revised 15 September 2020
Accepted 24 November 2020

[1]Department of Internal Medicine, University Center of Autoimmunity, University Medical Center Mainz, Mainz, Germany
[2]Monte Rosa Therapeutics, Basel, Switzerland
[3]Division of Rheumatology and Clinical Immunology, University Hospital Mainz, Mainz, Germany
[4]Acura Rheumatology Center Rhineland Palatinate, Bad Kreuznach, Germany

Correspondence to
Dr Sebastian Boegel;
seb.boegel@gmail.com

ABSTRACT

Objective Here, we assess the usage of high throughput sequencing (HTS) in rheumatic research and the availability of public HTS data of rheumatic samples.

Methods We performed a semiautomated literature review on PubMed, consisting of an R-script and manual curation as well as a manual search on the Sequence Read Archive for public available HTS data.

Results Of the 699 identified articles, rheumatoid arthritis (n=182 publications, 26%), systemic lupus erythematous (n=161, 23%) and osteoarthritis (n=152, 22%) are among the rheumatic diseases with the most reported use of HTS assays. The most represented assay is RNA-Seq (n=457, 65%) for the identification of biomarkers in blood or synovial tissue. We also find, that the quality of accompanying clinical characterisation of the sequenced patients differs dramatically and we propose a minimal set of clinical data necessary to accompany rheumatological-relevant HTS data.

Conclusion HTS allows the analysis of a broad spectrum of molecular features in many samples at the same time. It offers enormous potential in novel personalised diagnosis and treatment strategies for patients with rheumatic diseases. Being established in cancer research and in the field of Mendelian diseases, rheumatic diseases are about to become the third disease domain for HTS, especially the RNA-Seq assay. However, we need to start a discussion about reporting of clinical characterisation accompany rheumatological-relevant HTS data to make clinical meaningful use of this data.

Key messages

What is already known about this subject?

- High throughput sequencing (HTS) has enormous potential in rheumatic research as it offers a broad spectrum of molecular analysis.
- While widely adopted in cancer research, the usage of the various HTS assays in rheumatological research has not been quantified.

What does the study add?

- HTS is being adapted in rheumatological research, with rheumatoid arthritis and systemic lupus erythematous as the major indications and RNA-Seq as the most represented HTS assay.
- The quality of accompanying clinical characterisation of the sequenced patients differs dramatically.

How might this impact on clinical practice or future developments?

- Rheumatic diseases are about to become the third disease domain for HTS, however, here we start a discussion of reporting sequencing data by proposing a minimal set of clinical data necessary to accompany rheumatological-relevant HTS data.

INTRODUCTION

The aim of 'precision medicine' is the development of novel diagnosis, prevention and treatment strategies by taking into account the individuality of a patient [1] including the individual molecular profile.[2] The development of high throughput sequencing (HTS) platforms, collectively still called 'next-generation sequencing' (NGS), allows a comprehensive and multimodal molecular profile of a patient. In particular, gene expression analysis using whole- transcriptome sequencing (RNA-Seq) has become state-of-the-art [3] as it has been demonstrated to be more accurate, sensitive, as well as to have a broader dynamic range than DNA microarrays allowing the detection of more differentially expressed genes with higher fold change.[4] In addition, this assay provides both: abundance of transcripts and sequence information at base-pair resolution, thus allowing a broad spectrum of analyses beyond gene and transcript expression, enabling the detection of a wide variety of molecular features, such as alternative splicing events, RNA editing events, complementarity determining region 3 of T cell receptors (TCRs), B cell receptors (BCR), human leucocyte antigen (HLA) types.[5] In addition, HTS of exons, such as whole exome sequencing (WES) or targeted sequencing (gene panels), allows the rapid detection of DNA-encoded variants, such tumour cell mutations, and is a key technology enabling the development of mutanome-based cancer immunotherapies.[6] Not only has the adoption of HTS has been rapid in oncology, but clinical and research laboratories worldwide have made primary sequencing data available in the Sequence Read Archive (SRA, http://

www.ncbi.nlm.nih.gov/sra),[7] one of the largest data repositories with 7.5 PB of open-access HTS data.[8] The repository comprises data from over 340 000 samples [9] and thus provides a rich and valuable source for reanalysis of existing datasets with bioinformatic software [5] to identify novel and clinical translatable findings.

Moreover, non-invasive and minimally invasive profiling platforms, including 'liquid biopsies', allow one to obtain information about a disease state or response to treatment using, for example, blood from patients, followed by HTS profiling and subsequent bioinformatic analysis. While this concept is already implemented in oncology,[10] it is less mature in rheumatology. We argue here that HTS offers enormous potential to pave the way to personalised therapy [11] for patients with rheumatic diseases, particularly due to its extreme molecular and phenotypic heterogeneity [12].

Very recently in this journal, Kedra *et al* [13] reviewed the current use of big data and artificial intelligence in rheumatic diseases. Here, we focus on HTS profiling as a big data producer [14] and review both the literature using HTS and public HTS datasets in rheumatological diseases to quantify the adoption of this technology in rheumatology. In addition, we propose a minimal set of clinical data necessary to accompany rheumatological-relevant HTS data.

METHODS

Systematic literature review

The literature review was implemented in *R* (V.3.6.1,[15]) using the package easyPubMed (V.2.13,[16]) and consists of 2 steps. First an automated PubMed search was carried on 15 August 2020 out using the query string:

"(methylomics OR epigenomics OR NGS OR \"next generation sequencing\" OR RNA-Seq OR \"mRNA sequencing\" OR \"RNA sequencing\" OR \"RNA-sequencing\" OR \"transcriptome sequencing\" OR \"whole exome sequencing\" OR \"whole-exome sequencing\" OR \"high throughput sequencing\" OR \"high-throughput sequencing\" OR \"DNA sequencing\" OR \"RNA sequencing\" OR \"RNA-sequencing\" OR \"DNA-sequencing\" OR WXS OR WGS OR \"whole-genome sequencing\" OR \"whole genome sequencing\") AND (rheumatology OR \"rheumatologic disease\" OR \"rheumatologic disease\"))".

This search resulted in 1097 entries. The keywords of each returning dataset were intersected with official disease names extracted from International Statistical Classification of Diseases and Related Health Problems (ICD)-11[17] in order to filter out keywords that are not disease names. The remaining 253 keywords were then manually inspected to find rheumatic diseases. This approach identified the following diseases: autoinflammatory syndrome, dermatomyositis, enthesitis, familial mediterranean fever (FMF), granulomatosis with polyangiitis (GPA), juvenile idiopathic arthritis (JIA), myositis, osteoarthritis (OA), polymyositis, psoriatic arthritis (PsA), rheumatoid arthritis (RA), sacroiliitis, sjögren's syndrome, spondyloarthritis (SpA), synovitis, systemic lupus erythematosus (SLE), systemic sclerosis vasculitis, uveitis, gout and polychondritis.

In a second step more specific PubMed search was carried out using the disease names identified in the first step:

"(methylomics OR epigenomics OR NGS OR \"next generation sequencing\" OR RNA-Seq OR \"mRNA sequencing\" OR \"RNA sequencing\" OR \"RNA-sequencing\" OR \"transcriptome sequencing\" OR \"whole exome sequencing\" OR \"whole-exome sequencing\" OR \"high throughput sequencing\" OR \"high-throughput sequencing\" OR WXS OR WGS OR \"whole-genome sequencing\" OR \"whole genome sequencing\") AND (\"autoinflammatory syndrome\" OR dermatomyositis OR enthesitis OR \"familial mediterranean fever\" OR \"granulomatosis with polyangiitis\" OR \"juvenile idiopathic arthritis\" OR myositis OR osteoarthritis OR polymyositis OR \"psoriatic arthritis\" OR \"rheumatoid arthritis\" OR sacroiliitis OR \"sjögren syndrome\" OR \"sjögren's syndrome\" OR spondyloarthritis OR synovitis OR \"systemic lupus erythematosus\" OR \"systemic sclerosis\" OR vasculitis OR uveitis OR gout OR polychondritis)".

This search was carried on 4 September 2020 and resulted in 1162 PubMed hits, which were (if possible) annotated regarding disease name, PubMed ID, assay, journal, year of publication by automatic screening the title and abstract. Reviews (ie, publications which have 'Review' in metadata) and commentaries were excluded and missing information was added manually by manual inspection of the publication. After manual curation, 699 studies were included in this literature review (figure 1).

A list of all identified publications can be found at https://github.com/sebboegel/pubmed_rheuma_HTS.

SRA data analysis

Searching the SRA portal was carried out via the SRA portal at https://www.ncbi.nlm.nih.gov/sra using the diseases names identified in the literature review as key words one after another (ie, only one disease was searched at a time), then using the Run Selector ('Send results to Run Selector'), switching to the old Run Selector ('Revert to the old Run Selector') and downloading the metatable, which was input to a custom-built python script extracting all necessary information. In addition, the python package *pysradb* [18] was used for retrieving PubMed identifiers for an associated SRA project number.

Code availability

All scripts, input and result files, comments, as well all figures in this manuscript, generated with R package ggplot2 (V.3.2.1,[19]) are available at https://github.com/sebboegel/pubmed_rheuma_HTS.

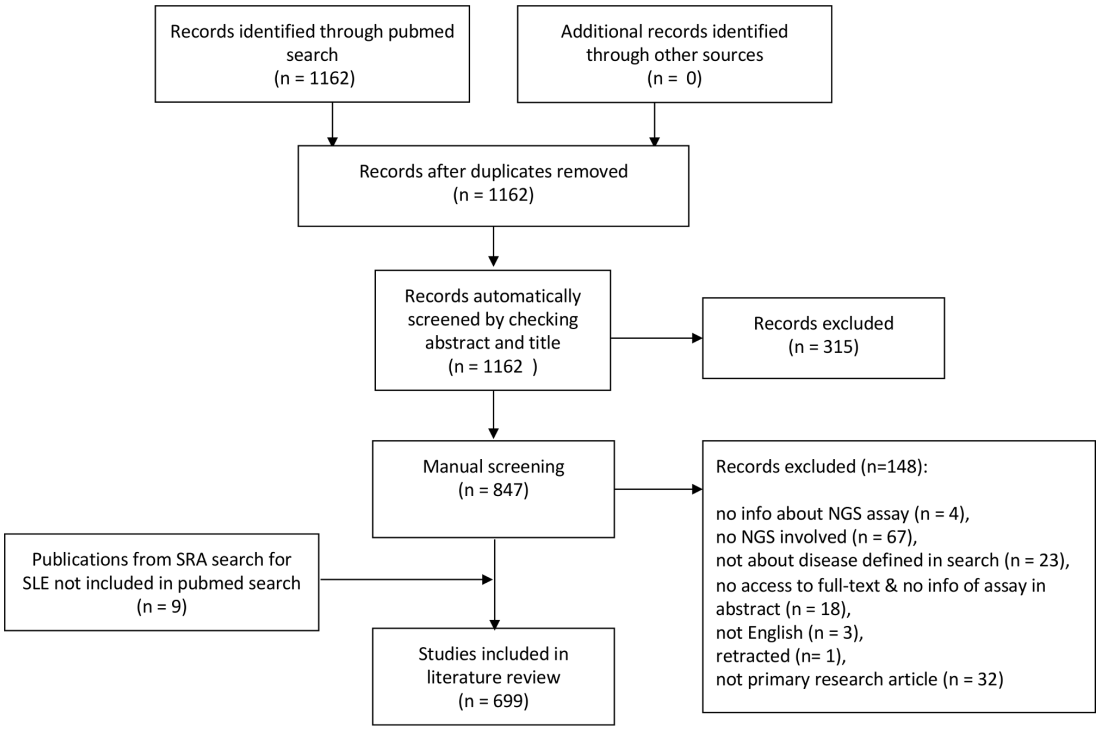

Figure 1 PRISMA flowdiagram of the literature review. For details, see the Methods section. NGS, next-generation sequencing; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; SLE, systemic lupus erythematous; SRA, Sequence Read Archive.

Paper counting

For counts that are not disease based (such as figure 2), the unique number of publications are depicted, which sum up to 699. However, as there exist publications using HTS on multiple rheumatic diseases, counting these papers in disease-based analysis (eg, figure 3) sum up to the total number of records (n=813), as a paper focusing on for example, SLE and RA will appear in the count for SLE and RA. Similarly, as there are publications using more than one HTS assay, summing up the number of assays discussed in the Results section will also exceed the number of unique publications.

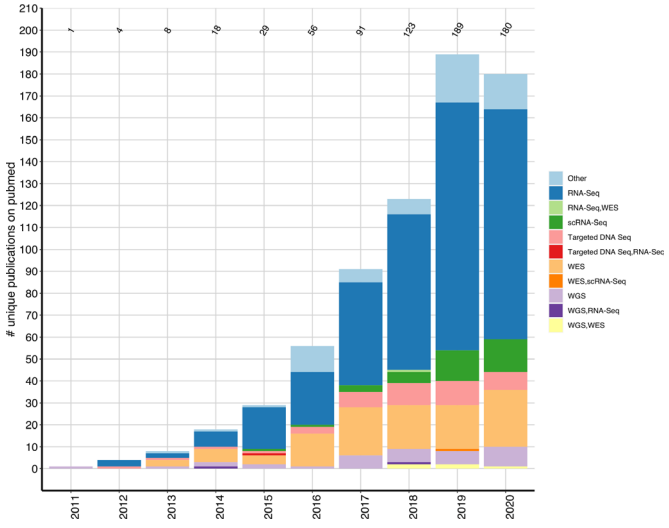

Figure 2 Publications per year. Number of unique identified primary research articles per year using different HTS assays in rheumatic diseases. HTS, high throughput sequencing; scRNA-Seq, single cell RNA-seq; WES, whole-exome sequencing; WGS, whole-genome sequencing.

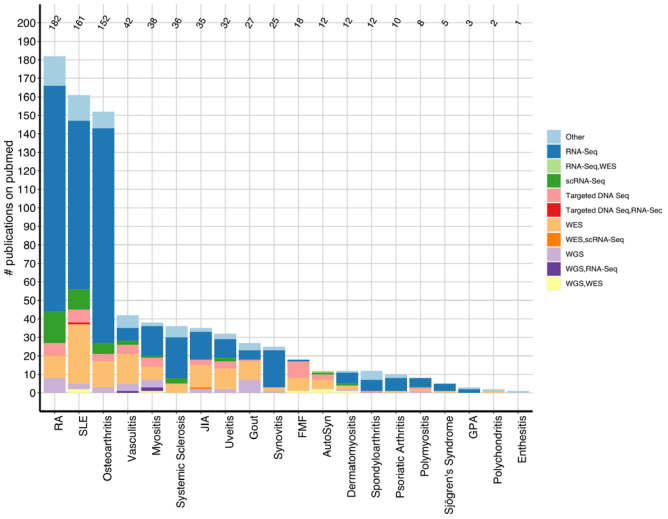

Figure 3 Publications and HTS assays per disease. Number of identified primary research publications per rheumatic disease using different HTS assays. AutoSyn, autoinflammatory syndrome; FMF, familial mediterranean fever; GPA, granulomatosis with polyangiitis; JIA, juvenile idiopathic arthritis; RA, rheumatoid arthritis; scRNA-Seq, single cell RNA-Seq; SLE, systemic lupus erythematosus; WES, whole-exome sequencing; WGS,whole-genome sequencing.

RESULTS

The semiautomated search strategy, consisting of an R-script and manual curation, resulted in 699 unique PubMed hits (813 total records). We analysed the identified literature according to the year of publication, the rheumatic diseases, the different HTS assays used, the wide variety of applications and the journals, in which these studies appeared.

HTS assays are adapted in rheumatic research: the number of papers including HTS published has increased from 18 in 2014 to 123 in 2018 and 189 in 2019 (figure 2). As of 4 September 2020, already 180 studies have been published and following this exponential growth, up to ~340 studies can be assumed by the end of 2020 (online supplemental figure S1). One of the first HTS studies we identified with this search strategy was published in 2011 and used whole-genome sequencing (WGS) to identify low-frequency variants associated in gout.[20 21]

RA, n=182/699 unique publications, 26%, SLE, n=161, 23% and OA (n=152, 22%) are the rheumatic diseases with the most reported use of HTS assays (figure 3). Applications of HTS in these diseases range from HLA typing,[22] TCR,[23 24] BCR,[25 26] and gene expression [27–29] profiling, as well as identification of T cell epitopes,[30] antibody repertoires,[31] and pathogenic mutations.[32 33]

The most represented assay is RNA-Seq (n=457, 65%) for the identification of biomarkers in blood or synovial tissue, for example, to distinguish active versus inactive/low disease activity states,[27] to examine response to anti-TNF therapy in RA,[34] to identify gene expression signatures correlating with disease phenotype,[35] for longitudinal analysis of peripheral blood TCR diversity in patients with SLE,[36] as well as for subgrouping patients with SLE with common clinical characteristics,[28] characterisation of circulating memory stem T cells in RA,[37] as well as to examine the BCR repertoire in patients with RA to identify B cell clones associated with autoreactivity.[38] In addition to messenger RNA, a wide range of RNA types can be measured, such as microRNAs (miRNAs) in RA,[39] JIA,[40] SLE and Sjögren's syndrome,[41] long non coding RNA (lncRNA) in SLE[42] as well as myositis,[43] and finally circular RNA as biomarker in SLE[44].

Transcriptomic analysis of individual cells (single cell RNA-Seq, scRNA-Seq) is increasingly becoming popular in cancer research,[45] for example, to better capture tumour heterogeneity. Here, we identify 40 out of the 457 RNA-Seq studies (9%, online supplemental figure S2) uses scRNA-Seq with applications in, for example, SLE for mapping disease heterogeneity at the single-cell level using the blood transcriptome[46] or for the identification of previously uncharacterised fibroblast subpopulations in the synovium of patients with RA.[47]

Applications for WES and targeted DNA (panel) sequencing (n=169, 24%) include identification of pathogenic mutations (mostly point mutations, small insertions and deletions) that can aid in diagnosis of monogenic autoinflammatory diseases and vasculitis,[48] FMF,[49] gout[50] or familial RA, SLE and primary Sjögren's syndrome[51] or Uveitis.[52] Of note, while HTS assays are powerful tools for large cohorts, we find many case reports using WES and gene panel sequencing in, for example, in a young patient with cutaneous vasculitis[53] or JIA,[54] as well as in a patient with RA experiencing immune dysregulation syndrome after abatacept therapy.[55]

We identified 42 (6%) studies using WGS. Again, the main application was identification of genetic variants, especially copy number variations, for example, of Fcγ receptor genes in RA[56] and association of mitochondrial genetic variation and copy number with gout,[57] as well as pharmacogenomic approaches examining patient's response to golimumab treatment explained by common single-nucleotide variations.[58]

Other assays (online supplemental figure S3) include the analysis of bacterial species using HTS (metagenomics, n=33; 5%) in, for example, a joint infection in a patient with SLE,[59] of the faecal microbiota of SLE mice[60] or the lung microbiota in early RA,[61] as well as epigenetic analysis (n=32, 5%) in SLE,[35 62] RA,[63 64] systemic sclerosis[65] and finally, phage immunoprecipitation sequencing (n=1) for HTS of autoantibody repertoires in systemic sclerosis.[66]

Among the journals in which these studies appeared, 'Arthritis and Rheumatology' (n=58, 8%), 'Annals of the Rheumatic Diseases' (n=40, 6%) and 'Plos One' (n=29, 4%) are the leading journals publishing papers covering a broad range of HTS assays, whereas the journals 'JCI Insight' (n=12, 2%) and 'Journal of Immunology' (n=9, 1%) focused so far on RNA-Seq, and 'Paediatric rheumatology online journal' (n=9) focus on WES (online supplemental figure S4) for the identification of disease-relevant genetic variants.

Raw-sequencing data in public domain

A search of samples in the SRA portal using the diseases identified in the PubMed search as key words revealed 17 023 HTS samples (figure 4) in 296 projects (online supplemental figure S5). The number of samples generated per study varies dramatically in the identified SRA projects (online supplemental figure S5) with 32/296 (11%) studies involving more than 100 study objects. Half of them (n=16) are produced in RA, seven in SLE, three in SpA, two in OA and JIA, and one in Systemic Sclerosis and GPA. The median number of HTS samples across the projects within the diseases is highest in GPA (72 samples/ study) and lowest in PsA 6.5 samples/ study). The vast majority of primary sequencing data originates from human biomaterial (15414/17023, 90.5%, online supplemental figure S6, primarily from samples reflecting the disease of interest (9854/17023, 58%, online supplemental figure S7, such as patients or disease models and 864 (5%) healthy controls. For the remaining 6305 (37%) samples, no phenotype or disease state was defined in the SRA metadata.

The majority of the samples are associated with RA (n=8483, 50%), SLE (n=3785, 22%) and OA (n=1386, 8%) and correlate with the relative abundance of studies

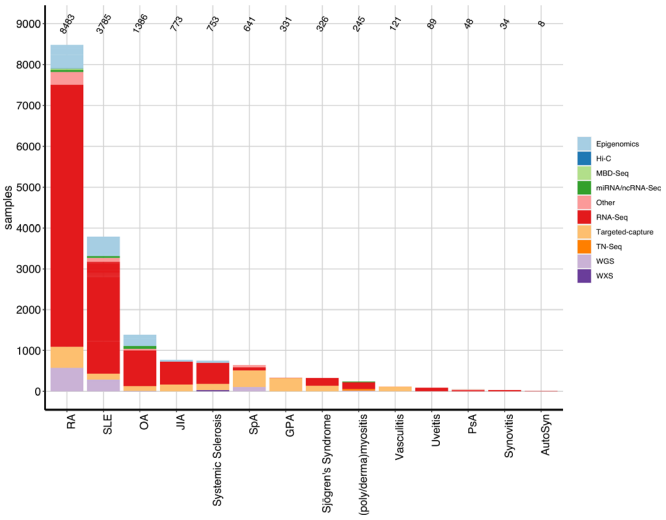

Figure 4 Public available high throughput sequencing (HTS) datasets. number of publicly available HTS samples on Sequence Read Archive for the rheumatic diseases identified in the literature review. AutoSyn, autoinflammatory syndrome; GPA, granulomatosis with polyangiitis; Hi-C, chromosome conformation capture; JIA, juvenile idiopathic arthritis; MBD-Seq, Methyl CpG binding domain-based capture and sequencing; miRNA-Seq, micro-RNA-Seq; ncRNA-Seq, non-coding-RNA-Seq; OA, osteoarthritis; PsA, psoriatic arthritis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; SpA, Spondyloarthritis; TN-Seq, transposon insertion sequencing; WGS, whole-genome sequencing; WXS, whole-exome sequencing.

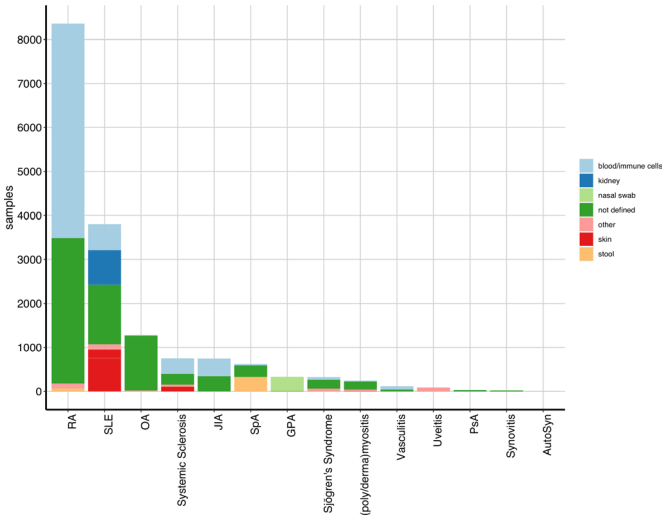

Figure 5 Tissue source of high throughput sequencing data on Sequence Read Archive (SRA). Distribution of tissues subject to sequencing in publicly available datasets on SRA. Disease abbreviations as in figure 4.

identified in the literature search for these diseases. Also, the dominance of the RNA-Seq assay is consistent with the PubMed findings. However, there are obvious inconsistencies when comparing the number of publications using or producing HTS data (figure 3) with the number of projects depositing HTS data on SRA (online supplemental figure S5). To examine this discrepancy, we used the RNA-Seq assay (including scRNA-Seq, miRNA, ncRNA) in SLE as an example for in depth analysis. By using the metadata table on the SRA website and a customised python script, followed by manual inspection, we identified 56 SRA projects, of which 43 projects provide raw RNA-Seq data. For seven of them no corresponding publication could be identified. Of the remaining 36 Projects, two SRP-IDs are associated with the same publication and two SRP-IDs are each associated with two different publications, resulting in 37 PubMed-IDs associated to SRA-Projects, which overlap with the 107 RNA-Seq studies in SLE identified in the PubMed search (figure 3). The remaining 70 publications were examined manually and 32/70 publications provided no information on the availability of the raw sequencing data at all, 13/70 provide the raw data 'on reasonable request', nine studies did not produce RNA-Seq data, but rather used publicly available datasets, six papers could not be accessed, three studies deposited the raw data at the European Genome Archive (EGA), two publications report an embargo on the data, that is, it will be provided with delay after the acceptance of the manuscript and one study made the data available under protected access at the database of Genotypes and Phenotypes (dbGaP) (online supplemental figure S8). Of note, four studies not providing the raw sequencing are case reports, which is consistent with FMF consisting primarily of case reports and we do not find any sequencing data from FMF on SRA (figure 4).

The most prominent sequencing platform is the Illumina HiSeq series (n=13 063, 77%, online supplemental figure S9) and paired end as preferred read layout (n=11 533, 68%), except for SLE with 1300 paired end and 2485 single end reads samples (online supplemental figure S10).

Analysing the tissue source of the HTS sample across different diseases (figure 5) reveals blood (whole blood, plasma, serum, peripheral blood mononuclear cell) and isolated immune cells (T cells, monocytes, dendritic cells) as the primary source material (6461/17 023, 38%). There are disease-specific preferences such as, cartilage in OA (84% of samples with defined tissue source), stool (faeces) in SpA, 87%, kidney in SLE (33%), synovium in PsA (52%) and synovitis (100%), as well as salivary gland in Sjögren's syndrome (50%), muscle in (poly/derma) myositis (63%) and retina in uveitis (100%) (figure 5 and online supplemental figure S11, 'sra_tissue.tsv').

Reporting of clinical patient data

Next, we examined the availability and quality of clinical information about the patients that were subject to sequencing and which HTS data is available from SRA. There are two challenges in finding patient characterisation of the primary HTS data of interest. First, the associated metadata does not use a defined ontology and no standardised patient/sample characterisation is required when deopsiting the sequencing data on SRA. Second challenge is the identification of the publication

Table 1 Examples of reporting clinical data for SLE patients subject to RNA-Seq

Category	Information provided	Bioproject	SRP-ID	Reference
No information	–	PRJNA431313	SRP131173	102
Rudimentary set	Demographics (age, sex, race)	PRJNA505280	SRP168421	103
Medium set of useful information	Demographics, serology, medication	PRJNA514365	SRP178271	104
	Demographics, disease duration, SLEDAI	PRJNA379992	SRP103040	105
Advanced set of useful information	Demographics, disease duration (years from diagnosis), SLEDAI, medication at time of blood draw (mg/day), ANA reactivity, prominent clinical features (such flares, new rash, etc), information for each patient individually.	PRJNA484966	SRP156584	106
	Demographics, SLEDAI, severity, immunosupprevise therapy, disease manifestation, comorbidities, medication, serology (eg, anti-DNA, C3/4)	PRJEB24742	SRP189104	107

Full list of patient characteristics available in file 'SLE_pubmed_rna_patient_info.csv' (see data availabiliy).
Bioproject: accessible at https://www.ncbi.nlm.nih.gov/bioproject/.
ANA, anti-nuclear antibody; C3/4, complement 3/4; SLE, systemic lupus erythematous; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index.

associated with the data. If no PubMed identifier is provided in the respective bioproject on SRA, the study can occasionally be identified by searching the bioproject title on PubMed or a related search engine.

In general, reporting of clinical data was highly diverse. In order to quantify this diversity, we used the RNA-Seq assay (including scRNA-Seq,miRNA, ncRNA) in SLE as an example for a detailed analysis. Of the 43 SRA projects providing SLE RNA-Seq data, 23 contain sequencing data from SLE patients, whereas the remaining projects deal with model organisms and cell lines (n=12) or the associated publications could be neither found (n=7) or accessed (n=1). Of these 23 projects, three associated manuscripts contain no information about the sequenced patient, four studies have at least a rudimentary set of information, eight publications with a medium set and eight papers with very detailed reporting of patient characteristics (table 1 and online supplemental figure S12).

DISCUSSION

High throughput gene expression profiling using DNA microarrays have already provided unprecedented views into the blood transcriptome of, for example, SLE,[67 68] RA,[69] SpA,[70] and thus paved the way for the development of personalised diagnostic and therapeutic strategies.

The introduction of 'next generation' HTS platforms, together with a tremendous evolution of open source bioinformatic software, enables the rapid detection of a wide variety of molecular features, such as alternative splicing events, RNA editing, HLA typing, BCR and TCR typing, mutation detection and many more,[5] thus adding new dimensions in understanding disease pathogenesis and biomarker identification.[71] Application and impact of HTS using NGS platforms in rheumatology have been reviewed in general[12] and for individual diseases, such as SLE[72] or RA.[73]

However, this is to our knowledge, the first study quantifying the usage of HTS in rheumatological research by reviewing literature on PubMed and examining public HTS data on SRA.

A limitation of this approach is that the numbers identified in this search are likely to be underestimated as potential publications may have been missed by the search. For example, one of the first studies using HTS for TCR and BCR repertoire analysis in RA was published April 2011[74] and is not indexed on PubMed (and thus has not been found by this search). Further, there exist more than 200 different rheumatic diseases[75] and our approach identified only a small subset (n=20). The strength of this approach is that it is easily reproducible. The provided R and python scripts along with all input and result files as well as comments about the manual steps of the analysis, enable reproduction of the results presented here and can be adopted for allowing literature review at any time point in the future.

A key finding is that HTS is indeed being adapted in rheumatological research with an exponential growth rate in number of publications since 2011. Major indications are RA and SLE, which are rheumatic disease with high prevalence rates of 0.5%–1% of the adult population in RA and 20–150 SLE cases per 100 000 individuals in the USA[76 77] in contrast to the many other rheumatic conditions that are classified as 'rare disease', such polymyositis (prevalence 1/14 000[78]). For the majority of the indications identified in this review, RNA-Seq was the most represented assay. While analysis of nucleotide variations by exome and genome sequencing holds great promise in the diagnosis of rare diseases,[79] going beyond the exome/genome, for example, analysing the gene expression to learn about pathomechanisms or personalised medicine approaches[12] results in the major challenge of very small patient populations.[80] Indeed, we find

that the majority of studies depositing sequencing data on the public repository SRA included low numbers of samples posing a challenge to the application of classical statistical analyses for target identification.[81] However, to be fair, not all projects we identified were designed to find biomarkers, such as case reports or mechanistic experiments using cell lines or model organisms.

The second key finding is that there exists a large number of raw sequencing data on the public repository SRA. However, we identified a gap between publications reporting usage of HTS assays and availability of this data on SRA. We quantified this gap with RNA-Seq projects for SLE as an example and found that the majority of studies not depositing data on SRA, do not provide any information about the availability of the primary sequencing data in the publication. Second most common finding was the information on the availability 'on request'. Reasons that might hinder researchers making HTS data publicly available might be technical or privacy challenges in sharing genomic data[82 83] or interests of the data owners.[84] With regard to privacy concerns, a feasible solution could be the deposition in repositories providing controlled and protected access to genomic data, such as the 'European Genome-Phenome Archive' (EGA)[85] or the 'database of Genoytpes and Phenotypes' (dbGaP).[86] EGA stores genomic data of 2953 studies[87] of which 1315 (45%) belong to 'cancer' and only 85 (3%) are labelled as 'Inflammatory' containing RA (n=19, 0.6%), SLE (n=7, 0.2%), ankylosing spondylitis (n=7) and psoriasis (n=1, 0.03%) datasets. As an example, very recently Panousis *et al* published a comprehensive genetic and transcriptomic profiling of 142 patients with SLE and 58 controls[27] and provided the raw and processed HTS data, clinical phenotypes/covariates, as well as the results of the genetic analysis under protected access (one needs to apply to access this data) at https://ega-archive.org/studies/EGAS00001003662. dbGaP is an online repository created by the National Center for Biotechnology Information provides controlled access to large-scale genomic datasets with associated phenotypes, such as 'The Cancer Genome Atlas' (TCGA)[88] or 'Genotype-Tissue Expression'.[89]

Sharing HTS data have several advantages. First of all, when data are made available for reuse, citations to the initial report increase.[90] In addition, genomic data potentially has value beyond the initial purpose and re-analysis of publicly available sequencing data with novel bioinformatic tools can lead to novel insights, for example, in RA,[91] to examine HLA and proteasome expression in different tissues[92] or public HTS data can be used to provide supportive information in addition to own sequencing experiments, as in the case of uncovering distinct subsets of patients with SLE using machine learning methods.[93] However, clinically useful and translational reanalysis requires (1) the searchability of this data, which is only guaranteed if the data are deposited one of the above-mentioned repositories and (2) the availability of detailed patient characteristics along with clinical information linked to the respective sequencing sample (ie, data characterisation challenge).[94]

Very recently, Gossec *et al* present 10 EULAR points to consider (PTC) for the use of big data, including 'omics and imaging data, in rheumatic and musculoskeletal diseases.[95] Here, we emphasise the importance of clinical data linked to the patient HTS data and propose an additional PTC: 'provide clinical characterisation'. It is necessary to agree on a set of rules for reporting clinical data in the context of genomic sequencing experiments, link them to the respective sequencing sample of the patient to connect genotype (eg, genome) with phenotype (eg, treatment response, organ manifestation, grade of disease) and extract as much clinically translatable information as possible from existing data. A successful example from cancer research is TCGA, which is a cancer genomics programme consisting of research centres worldwide, generating genomic, epigenomic, transcriptomic and proteomic data of more than 30 cancer types including histopathological images and clinical data. To make clinically valuable analysis comparable between the projects within the consortium, such as survival outcome analysis[96] guidelines on reporting clinical data were developed[97] and a data dictionary was defined to define necessary clinical entities, such as 'Demographic', 'Diagnosis', 'Family History', 'Treatment' and 'Follow-up'.[98] We recognise that there is rheumatic disease specific information that is important to share, for example, Schirmer test for Sjögren syndrome. Nevertheless, we translate these guidelines into the world of rheumatology

Table 2 Proposal of a minimal set of clinical information when sharing patient HTS data to enable clinically useful reanalysis

Clinical entity	Data points
Demographic	ethnicity, gender, age
Diagnosis	Primary diagnosis (ICD-10), type/grade/stage, disease activity scores (SLEDAI, BASFI, BASDAI, VAS, DAS, …)
Exposure	cigarettes_per_day, years smoked
Family history	History of autoimmune disease in family
Follow-up	BMI, comorbidities, progression or recurrence, weight, disease duration
Molecular tests	Anti-CCP, HLA status, C3, C4, autoantibodies (ANA, ENA…)
Treatment	Therapeutic agents, dose, frequency, outcome, adverse events

ANA, anti-nuclear antibody; BASDAI, Bath Ankylosing Spondylitis Disease Activity Index; BASFI, Bath Ankylosing Spondylitis Functional Index; C3/4, Complement 3/4; CCP, Citrullinated Peptide/Protein antibodies; DAS, Diseases Activity Score; ENA, Extractable Nuclear Antigen; HLA, Human Leukocyte Antigen; HTS, high throughput sequencing; ICD, International Statistical Classification of Diseases and Related Health Problems; SLEDAI, Systemic Lupus Erythematosus Disease Activity Index; VAS, Visual Analog Scale.

and propose a minimal set of clinical data to be reported in HTS experiments (table 2).

Being already established in cancer research and in the field of Mendelian diseases,[99] rheumatic diseases are about to become the third disease domain for HTS. This is an important observation, as many of the bioinformatic tools for analysing HTS data have been developed in the context of cancer research. Not all of them can be directly applied to rheumatology, such as mutation detection tools, and require adoption to rheumatological datasets. We foresee an evolution of bioinformatic software newly developed or adopted to the specific needs and questions of rheumatological disease. Especially the RNA-Seq assay, which we found already widely adopted in rheumatology, will be a central and powerful assay in deciphering pathomechanisms, precision approaches and might lead to new disease definitions based on molecular characteristics as it has been shown in cancer.[100] However, there is a need for a global solution for sharing clinical and genomic data.[101] This discussion started in cancer research and must continue in rheumatic research.

Twitter Sebastian Boegel @sebboeg

Acknowledgements We wish to thank the German Network for Bioinformatics Infrastructure (de.NBI) for providing the compute infrastructure for this analysis on the de.NBI Cloud. We are grateful to the anonymous reviewers for significantly improving this manuscript.

Contributors SB wrote the R and python scripts and performed the analysis. SB, JCC and AS wrote the manuscript. AS supervised the study. All authors reviewed the manuscript and provided comments.

Funding We wish to acknowledge the RARENET EU-Interreg for supporting this study. Furthermore, SB wishes to acknowledge and thank the Mainz Research School of Translational Biomedicine (TransMed) for support.

Competing interests SB and JCC have nothing to declare. AS has received speaker fees (less than US$10 000) and grant/research support by AbbVie, Novartis, Roche and GSK.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data relevant to the study and the R script are included in the article or available on at https://github.com/sebboegel/pubmed_rheuma_HTS.

ORCID iD

Sebastian Boegel http://orcid.org/0000-0001-5425-8304

REFERENCES

1 Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–5.
2 Ashley EA. Towards precision medicine. *Nat Rev Genet* 2016;17:507–22.
3 Byron SA, Van Keuren-Jensen KR, Engelthaler DM, *et al*. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* 2016;17:257–71.
4 Zhao S, Fung-Leung W-P, Bittner A, *et al*. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014;9:e78644.
5 Boegel S, Castle JC, Kodysh J, *et al*. Bioinformatic methods for cancer neoantigen prediction. *Prog Mol Biol Transl Sci* 2019;164:25–60.
6 Vormehr M, Diken M, Boegel S, *et al*. Mutanome directed cancer immunotherapy. *Curr Opin Immunol* 2016;39:14–22.
7 Leinonen R, Sugawara H, Shumway M, *et al*. The sequence read archive. *Nucleic Acids Res* 2011;39:D19–21.
8 SRA statistics. Available: https://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi [Accessed 25 Mar 2020].
9 SRA - NCBI, 2020. Available: https://www.ncbi.nlm.nih.gov/sra/?term=cancer [Accessed 31 Mar 2020].
10 Karachaliou N, Mayo-de-Las-Casas C, Molina-Vila MA, *et al*. Real-time liquid biopsies become a reality in cancer treatment. *Ann Transl Med* 2015;3:36.
11 Isaacs JD, Ferraccioli G. The need for personalised medicine for rheumatoid arthritis. *Ann Rheum Dis* 2011;70:4–7.
12 Donlin LT, Park S-H, Giannopoulou E, *et al*. Insights into rheumatic diseases from next-generation sequencing. *Nat Rev Rheumatol* 2019;15:327–39.
13 Kedra J, Radstake T, Pandit A, *et al*. Current status of use of big data and artificial intelligence in RMDs: a systematic literature review informing EULAR recommendations. *RMD Open* 2019;5:e001004.
14 Stephens ZD, Lee SY, Faghri F, *et al*. Big data: astronomical or Genomical? *PLoS Biol* 2015;13:e1002195.
15 Website. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019. https://www.R-project.org/
16 Damiano Fantini. easyPubMed: search and Retrieve scientific publication records from PubMed. R package version 2.13, 2019. Available: https://CRAN.R-project.org/package=easyPubMed [Accessed 1 Apr 2020].
17 ICD-11. Available: https://icd.who.int/dev11/f/en [Accessed 20 Feb 2020].
18 Choudhary S. pysradb: a python package to query next-generation sequencing metadata and data from NCBI sequence read Archive. *F1000Res* 2019;8:532.
19 Wickham H. Programming with ggplot2. *Use R!* 2016:241–53.
20 Sulem P, Gudbjartsson DF, Walters GB, *et al*. Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet* 2011;43:1127–30.
21 Bakir-Gungor B, Sezerman OU. A new methodology to associate SNPs with human diseases according to their pathway related context. *PLoS One* 2011;6:e26277.
22 Huang Y-H, Khor S-S, Zheng X. A high-resolution HLA imputation system for the Taiwanese population: a study of the Taiwan Biobank. *Pharmacogenomics J*.
23 Musters A, Klarenbeek PL, Doorenspleet ME, *et al*. In rheumatoid arthritis, synovitis at different inflammatory sites is dominated by shared but patient-specific T cell clones. *J Immunol* 2018;201:417–22.
24 Sakurai K, Ishigaki K, Shoda H, *et al*. Hla-Drb1 shared epitope alleles and disease activity are correlated with reduced T cell receptor repertoire diversity in CD4+ T cells in rheumatoid arthritis. *J Rheumatol* 2018;45:905–14.
25 Shi X, Shao T, Huo F, *et al*. An analysis of abnormalities in the B cell receptor repertoire in patients with systemic sclerosis using high-throughput sequencing. *PeerJ* 2020;8:e8370.
26 Wang Y, Lloyd KA, Melas I, *et al*. Rheumatoid arthritis patients display B-cell dysregulation already in the naïve repertoire consistent with defects in B-cell tolerance. *Sci Rep* 2019;9:19995.
27 Panousis NI, Bertsias GK, Ongen H, *et al*. Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. *Ann Rheum Dis* 2019;78:1079–89.
28 Rai R, Chauhan SK, Singh VV, *et al*. Rna-Seq analysis reveals unique transcriptome signatures in systemic lupus erythematosus patients with distinct autoantibody specificities. *PLoS One* 2016;11:e0166312.
29 Mittal A, Pachter L, Nelson JL, *et al*. Pregnancy-Induced changes in systemic gene expression among healthy women and women with rheumatoid arthritis. *PLoS One* 2015;10:e0145204.
30 Liu M, Degner J, Davis JW, *et al*. Identification of HLA-DRB1 association to adalimumab immunogenicity. *PLoS One* 2018;13:e0195325.
31 Kinslow JD, Blum LK, Deane KD, *et al*. Elevated IgA Plasmablast levels in subjects at risk of developing rheumatoid arthritis. *Arthritis Rheumatol* 2016;68:2372–83.
32 Ahmad M, Hermanson ME, Enzenauer R, *et al*. Lipogranulomatous subconjunctival nodules: a novel presentation in Blau syndrome. *J Aapos* 2017;21:249–51.
33 Vahidnezhad H, Youssefian L, Sotoudeh S, *et al*. Genomics-based treatment in a patient with two overlapping heritable skin disorders:

epidermolysis bullosa and acrodermatitis enteropathica. *Hum Mutat* 2020;41:906–12.
34 Farutin V, Prod'homme T, McConnell K, *et al*. Molecular profiling of rheumatoid arthritis patients reveals an association between innate and adaptive cell populations and response to anti-tumor necrosis factor. *Arthritis Res Ther* 2019;21:216.
35 Zhao M, Liu S, Luo S, *et al*. DNA methylation and mRNA and microRNA expression of SLE CD4+ T cells correlate with disease phenotype. *J Autoimmun* 2014;54:127–36.
36 Thapa DR, Tonikian R, Sun C, *et al*. Longitudinal analysis of peripheral blood T cell receptor diversity in patients with systemic lupus erythematosus by next-generation sequencing. *Arthritis Res Ther* 2015;17:132.
37 Cianciotti BC, Ruggiero E, Campochiaro C, *et al*. Cd4+ memory stem T cells recognizing citrullinated epitopes are expanded in patients with rheumatoid arthritis and sensitive to tumor necrosis factor blockade. *Arthritis Rheumatol* 2020;72:565–75.
38 Doorenspleet ME, Klarenbeek PL, de Hair MJH, *et al*. Rheumatoid arthritis synovial tissue harbours dominant B-cell and plasma-cell clones associated with autoreactivity. *Ann Rheum Dis* 2014;73:756–62.
39 Dunaeva M, Blom J, Thurlings R, *et al*. Circulating serum miR-223-3p and miR-16-5p as possible biomarkers of early rheumatoid arthritis. *Clin Exp Immunol* 2018;193:376–85.
40 Nziza N, Jeziorski E, Delpont M, *et al*. Synovial-Fluid miRNA signature for diagnosis of juvenile idiopathic arthritis. *Cells* 2019;8. doi:10.3390/cells8121521. [Epub ahead of print: 26 Nov 2019].
41 Chen J-Q, Papp G, Póliska S, *et al*. Microrna expression profiles identify disease-specific alterations in systemic lupus erythematosus and primary Sjögren's syndrome. *PLoS One* 2017;12:e0174585.
42 Ye H, Wang X, Wang L, *et al*. Full high-throughput sequencing analysis of differences in expression profiles of long noncoding RNAs and their mechanisms of action in systemic lupus erythematosus. *Arthritis Res Ther* 2019;21:70.
43 Hamann PD, Roux BT, Heward JA, *et al*. Transcriptional profiling identifies differential expression of long non-coding RNAs in Jo-1 associated and inclusion body myositis. *Sci Rep* 2017;7:8024.
44 Guo G, Wang H, Ye L, *et al*. Hsa_circ_0000479 as a novel diagnostic biomarker of systemic lupus erythematosus. *Front Immunol* 2019;10:2281.
45 Zhang X, Marjani SL, Hu Z, *et al*. Single-Cell sequencing for precise cancer research: progress and prospects. *Cancer Res* 2016;76:1305–12.
46 Nehar-Belaid D, Hong S, Marches R, *et al*. Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nat Immunol* 2020;21:1094–106.
47 Stephenson W, Donlin LT, Butler A, *et al*. Single-Cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nat Commun* 2018;9:791.
48 Omoyinmi E, Standing A, Keylock A, *et al*. Clinical impact of a targeted next-generation sequencing gene panel for autoinflammation and vasculitis. *PLoS One* 2017;12:e0181874.
49 Terhaar C, Teed N, Allen R, *et al*. Clinical experience with multigene carrier panels in the reproductive setting. *Prenat Diagn* 2018. doi:10.1002/pd.5272. [Epub ahead of print: 23 Apr 2018].
50 Huang X-F, Sun L, Zhang C, *et al*. Whole-Exome Sequencing Reveals a Rare Missense Variant in *SLC16A9* in a Pedigree with Early-Onset Gout. *Biomed Res Int* 2020;2020:4321419.
51 Wang Y, Chen S, Chen J, *et al*. Germline genetic patterns underlying familial rheumatoid arthritis, systemic lupus erythematosus and primary Sjögren's syndrome highlight T cell-initiated autoimmunity. *Ann Rheum Dis* 2020;79:268–75.
52 Zerkaoui M, Laarabi FZ, Ajhoun Y, *et al*. A novel single variant in the MEFV gene causing Mediterranean fever and Behçet's disease: a case report. *J Med Case Rep* 2018;12:53.
53 Mauro A, Omoyinmi E, Sebire NJ, *et al*. De Novo *PTEN* Mutation in a Young Boy with Cutaneous Vasculitis. *Case Rep Pediatr* 2017;2017:9682803.
54 AlKhater S. A novel mutation in *NCF2* resulting in very-early-onset colitis and juvenile idiopathic arthritis in a patient with chronic granulomatous disease. *Allergy Asthma Clin Immunol* 2019;15:68.
55 Ureshino H, Koarada S, Kamachi K, *et al*. Immune dysregulation syndrome with de novo CTLA4 germline mutation responsive to abatacept therapy. *Int J Hematol* 2020;111:897–902.
56 Franke L, el Bannoudi H, Jansen DTSL, *et al*. Association analysis of copy numbers of FC-gamma receptor genes for rheumatoid arthritis and other immune-mediated phenotypes. *Eur J Hum Genet* 2016;24:263–70.
57 Gosling AL, Boocock J, Dalbeth N, *et al*. Mitochondrial genetic variation and gout in Māori and Pacific people living in Aotearoa New Zealand. *Ann Rheum Dis* 2018;77:571–8.
58 Standish KA, Huang CC, Curran ME, *et al*. Comprehensive analysis of treatment response phenotypes in rheumatoid arthritis for pharmacogenetic studies. *Arthritis Res Ther* 2017;19:90.
59 Huang Y, Ma Y, Miao Q, *et al*. Arthritis caused by *Legionella micdadei* and *Staphylococcus aureus*: metagenomic next-generation sequencing provides a rapid and accurate access to diagnosis and surveillance. *Ann Transl Med* 2019;7:589.
60 Ma Y, Xu X, Li M, *et al*. Gut microbiota promote the inflammatory response in the pathogenesis of systemic lupus erythematosus. *Mol Med* 2019;25:35.
61 Scher JU, Joshua V, Artacho A, *et al*. The lung microbiota in early rheumatoid arthritis and autoimmunity. *Microbiome* 2016;4:60.
62 Scharer CD, Blalock EL, Barwick BG, *et al*. Atac-Seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells. *Sci Rep* 2016;6:27030.
63 Ham S, Bae J-B, Lee S, *et al*. Epigenetic analysis in rheumatoid arthritis synoviocytes. *Exp Mol Med* 2019;51:1–13.
64 Loh C, Park S-H, Lee A, *et al*. TNF-Induced inflammatory genes escape repression in fibroblast-like synoviocytes: transcriptomic and epigenomic analysis. *Ann Rheum Dis* 2019;78:1205–14.
65 Tsou P-S, Wren JD, Amin MA, *et al*. Histone deacetylase 5 is overexpressed in scleroderma endothelial cells and impairs angiogenesis via repression of proangiogenic factors. *Arthritis Rheumatol* 2016;68:2975–85.
66 Xu GJ, Shah AA, Li MZ, *et al*. Systematic autoantigen analysis identifies a distinct subtype of scleroderma with coincident cancer. *Proc Natl Acad Sci U S A* 2016;113:E7526–34.
67 Banchereau R, Hong S, Cantarel B, *et al*. Personalized Immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 2016;165:551–65.
68 Haynes WA, Haddon DJ, Diep VK, *et al*. Integrated, multicohort analysis reveals unified signature of systemic lupus erythematosus. *JCI Insight* 2020;5. doi:10.1172/jci.insight.122312. [Epub ahead of print: 27 Feb 2020].
69 Macías-Segura N, Castañeda-Delgado JE, Bastian Y, *et al*. Transcriptional signature associated with early rheumatoid arthritis and healthy individuals at high risk to develop the disease. *PLoS One* 2018;13:e0194205.
70 Gu J, Märker-Hermann E, Baeten D, *et al*. A 588-gene microarray analysis of the peripheral blood mononuclear cells of spondyloarthropathy patients. *Rheumatology* 2002;41:759–66.
71 Robinson WH, Mao R. Biomarkers to guide clinical therapeutics in rheumatology? *Curr Opin Rheumatol* 2016;28:168–75.
72 Rai G, Rai R, Saeidian AH, *et al*. Microarray to deep sequencing: transcriptome and miRNA profiling to elucidate molecular pathways in systemic lupus erythematosus. *Immunol Res* 2016;64:14–24.
73 Sumitomo S, Nagafuchi Y, Tsuchida Y, *et al*. Transcriptome analysis of peripheral blood from patients with rheumatoid arthritis: a systematic review. *Inflamm Regen* 2018;38:21.
74 Klarenbeek PL, Doorenspleet ME, van Schaik BDC, *et al*. Complete T and B cell receptor repertoire analysis in rheumatoid arthritis using high throughput sequencing. *Ann Rheum Dis* 2010;69:A33–4.
75 EULAR. Available: https://www.eular.org/myUploadData/files/10%20things%20on%20RD.pdf [Accessed 2 Apr 2020].
76 Lawrence RC, Helmick CG, Arnett FC, *et al*. Estimates of the prevalence of arthritis and selected musculoskeletal disorders in the United States. *Arthritis Rheum* 1998;41:778–99.
77 Chakravarty EF, Bush TM, Manzi S, *et al*. Prevalence of adult systemic lupus erythematosus in California and Pennsylvania in 2000: estimates obtained using hospitalization data. *Arthritis Rheum* 2007;56:2092–4.
78 Inserm Us14- All. Orphanet: search a disease. Available: https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=700 [Accessed 2 Apr 2020].
79 Liu Z, Zhu L, Roberts R, *et al*. Toward clinical implementation of next-generation sequencing-based genetic testing in rare diseases: where are we? *Trends Genet* 2019;35:852–67.
80 Frésard L, Montgomery SB. Diagnosing rare diseases after the exome. *Cold Spring Harb Mol Case Stud* 2018;4. doi:10.1101/mcs.a003392. [Epub ahead of print: 17 Dec 2018].
81 Li C-I, Samuels DC, Zhao Y-Y, *et al*. Power and sample size calculations for high-throughput sequencing-based experiments. *Brief Bioinform* 2018;19:1247–55.
82 Corpas M, Kovalevskaya NV, McMurray A, *et al*. A fair guide for data providers to maximise sharing of human genomic data. *PLoS Comput Biol* 2018;14:e1005873.
83 Brown AV, Campbell JD, Assefa T, *et al*. Ten quick tips for sharing open genomic data. *PLoS Comput Biol* 2018;14:e1006472.

84 Nanda S, Kowalczuk MK. Unpublished genomic data-how to share? *BMC Genomics* 2014;15:5–2.
85 Lappalainen I, Almeida-King J, Kumanduri V, *et al*. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* 2015;47:692–5.
86 Wong KM, Langlais K, Tobias GS, *et al*. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res* 2017;45:D819–26.
87 European Genome-phenome Archive. Available: https://ega-archive.org [Accessed 2 Apr 2020].
88 Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, *et al*. The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;45:1113–20.
89 Melé M, Ferreira PG, Reverter F, *et al*. Human genomics. the human transcriptome across tissues and individuals. *Science* 2015;348:660–5.
90 Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. *PeerJ* 2013;1:e175.
91 Platzer A, Nussbaumer T, Karonitsch T, *et al*. Analysis of gene expression in rheumatoid arthritis and related conditions offers insights into sex-bias, gene biotypes and co-expression patterns. *PLoS One* 2019;14:e0219698.
92 Boegel S, Löwer M, Bukur T, *et al*. Hla and proteasome expression body MAP. *BMC Med Genomics* 2018;11:36.
93 Figgett WA, Monaghan K, Ng M, *et al*. Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. *Clin Transl Immunology* 2019;8:e01093.
94 Learned K, Durbin A, Currie R, *et al*. Barriers to accessing public cancer genomic data. *Sci Data* 2019;6:1–7.
95 Gossec L, Kedra J, Servy H, *et al*. EULAR points to consider for the use of big data in rheumatic and musculoskeletal diseases. *Ann Rheum Dis* 2020;79:69–76.
96 Liu J, Lichtenberg T, Hoadley KA, *et al*. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 2018;173:400–16. e11.
97 Clinical data harmonization | NCI genomic data commons. Available: https://gdc.cancer.gov/about-data/data-harmonization-and-generation/clinical-data-harmonization [Accessed 2 Apr 2020].
98 Viewer - GDC Docs. Available: https://docs.gdc.cancer.gov/Data_Dictionary/viewer/#?view=table-entity-list&anchor=clinical [Accessed 2 Apr 2020].
99 Jamuar SS, Tan E-C. Clinical application of next-generation sequencing for Mendelian diseases. *Hum Genomics* 2015;9:10.
100 Hoadley KA, Yau C, Wolf DM, *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014;158:929–44.
101 Clinical Cancer Genome Task Team of the Global Alliance for Genomics and Health, Lawler M, Haussler D, *et al*. Sharing Clinical and Genomic Data on Cancer - The Need for Global Solutions. *N Engl J Med* 2017;376:2006–9.
102 Quero L, Tiaden AN, Hanser E, *et al*. miR-221-3p drives the shift of M2-Macrophages to a pro-inflammatory function by suppressing JAK3/STAT3 activation. *Front Immunol* 2019;10:3087.
103 Tokuyama M, Kong Y, Song E, *et al*. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci U S A* 2018;115:12565–72.
104 Sirobhushanam S, Parsa N, Reed TJ, *et al*. Staphylococcus aureus colonization is increased on lupus skin lesions and is promoted by IFN-mediated barrier disruption. *J Invest Dermatol* 2020;140:1066–74.
105 Der E, Ranabothu S, Suryawanshi H, *et al*. Single cell RNA sequencing to dissect the molecular heterogeneity in lupus nephritis. *JCI Insight* 2017;2. doi:10.1172/jci.insight.93009. [Epub ahead of print: 04 May 2017].
106 Scharer CD, Blalock EL, Mi T, *et al*. Epigenetic programming underpins B cell dysfunction in human SLE. *Nat Immunol* 2019;20:1071–82.
107 Greiling TM, Dehner C, Chen X, *et al*. Commensal orthologs of the human autoantigen Ro60 as triggers of autoimmunity in lupus. *Sci Transl Med* 2018;10. doi:10.1126/scitranslmed.aan2306. [Epub ahead of print: 28 Mar 2018].