

SCIENTIFIC REPORTS



OPEN

Genetic insights into juvenile idiopathic arthritis derived from deep whole genome sequencing

Laiping Wong¹, Kaiyu Jiang¹, Yanmin Chen¹ & James N. Jarvis^{1,2}

Deep whole genome sequencing (WGS) allows for the comprehensive study of genetic landscapes at finer resolution than array based methods. We conducted deep WGS on children with the polyarticular form of juvenile idiopathic arthritis (JIA), using 2 independent cohorts to ascertain the sequencing fidelity. Genome wide SNP density analysis identified 18 SNP hotspots with comparison to the 1000 Genome Projects (1KGP) data. A subset of the genes adjacent to SNP hotspots showed statistically significant enrichment in immunological processes. Genes adjacent to indel hotspots were functionally related to G-protein coupled signaling pathways. Further analyses elucidated significantly more JIA SNPs with regulatory potential compared to 1KGP data. Furthermore, SNPs located within linkage disequilibrium (LD) blocks containing previously identified JIA-associated SNPs demonstrated higher regulation potential compared to SNPs outside LD blocks. We also demonstrated enrichment of novel JIA variants in histone modification peaks and DNase hypersensitivity sites in B cells. This study greatly expands the number of genetic variants that may contribute to JIA and give us some clues into what may trigger this disease. To date, this study is the first deep WGS effort on children with JIA and provides useful genetic resources for research communities particularly in understanding JIA etiology.

JIA is an archetypal complex trait, in which small but measurable genetic susceptibility can be identified in multiple genetic loci, and where the environment contributes significant risk¹. In JIA, the search for causal variants that contribute to disease risk has been hampered by the small effect sizes². Furthermore, although multiple risk loci for JIA have been identified via candidate gene approaches and genetic fine mapping studies^{3,4}, these approaches carry an inherent bias in their assumption that JIA is an “autoimmune disease”. Genome-wide association studies are free from inherent bias about disease pathogenesis⁵, but these studies provide only a rough estimate of genetic risk loci and are limited in their capacity to detect regions conferring subtle but biologically significant risk. Finally, the task of finding causal variants is further complicated by the fact that known genetic risk for JIA, resides primarily within the non-coding genome⁶.

Whole genome sequencing (WGS), with sufficient depth, may provide, in an unbiased manner, a new opportunity to obtain a finer resolution for regions of genetic risk in complex traits, identify new regions not identified through chip-based methods that query targeted regions on assumptions regarding underlying pathogenic mechanisms. To explore this possibility, we performed WGS on children with polyarticular JIA. We report here multiple new genetic variations and identify epigenetic landscapes surrounding these genetic mutations that may cast light on previously unrecognized disease mechanisms.

Results

Reproducibility of genetic variants discovery. As our sequencing data were obtained through two independent sequencing experiments, we conducted a variants concordance check for the SNPs and indels identified in these two sets of data (biological replication). We define concordance rate as a ratio of the number of shared variants between two sets to the total number of variants in the observed set. We observed averages of 77.14% and 66.63% concordance rates for SNPs and indels respectively (Supplementary Figure 1). When we examined genetic variation of a pair of biological replicates from the first batch (technical replication), we saw concordance rates of 99% and 83% for SNPs and indels respectively (Supplementary Table 1). Given these high concordance

¹Department of Pediatrics, University at Buffalo, Buffalo, NY, 14203, USA. ²Genetics, Genomics, & Bioinformatics Program, University at Buffalo, Buffalo, NY, 14203, USA. Correspondence and requests for materials should be addressed to J.N.J. (email: jamesjar@buffalo.edu)

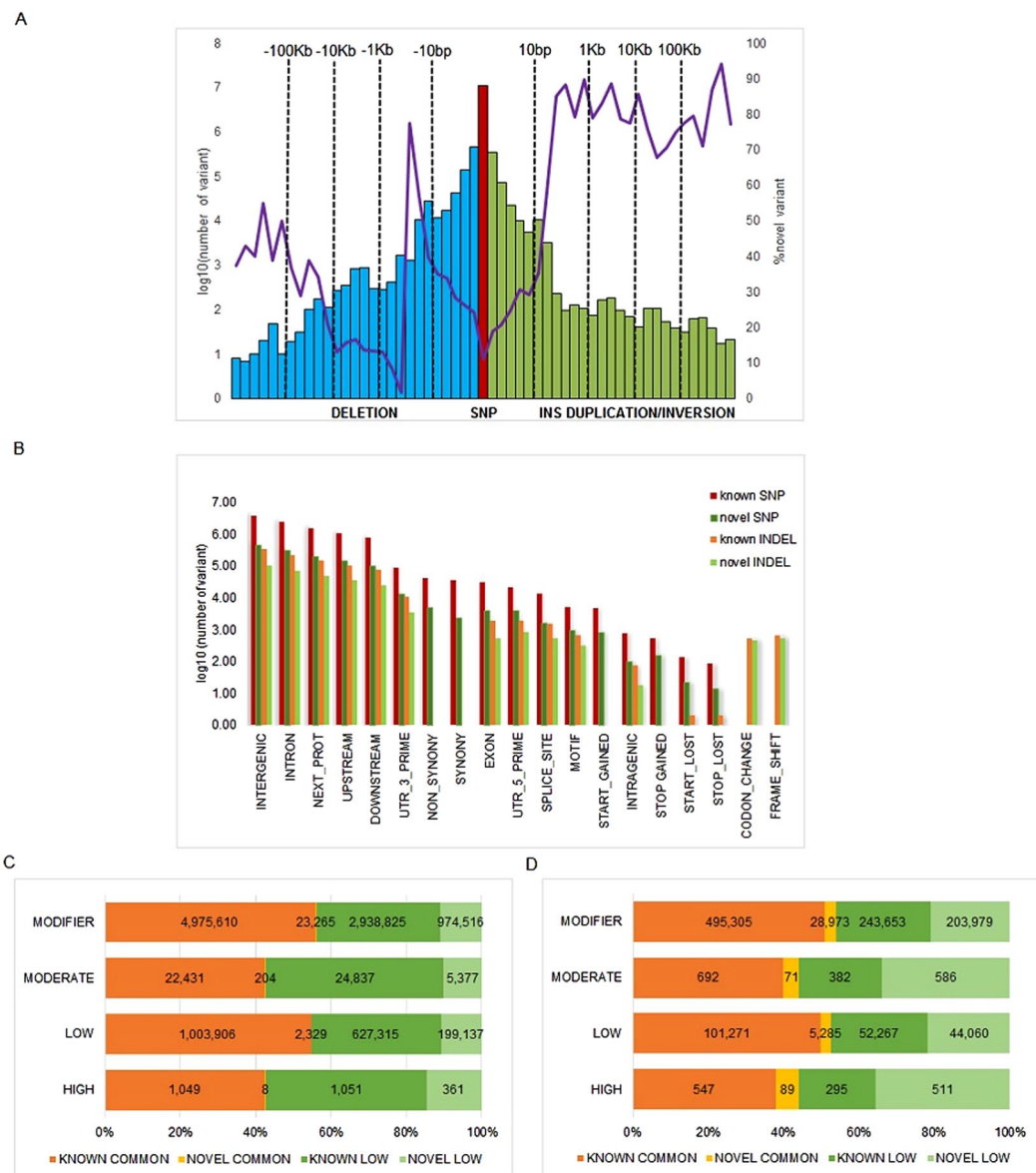


Figure 1. Variants discovered from high coverage DNA whole genome sequencing on 48 JIA individuals. **(A)** Bars show variant numbers (\log_{10}) stratified by sizes (horizontal axis) for different type of variant, SNPs are in brown, indels consisting of insertions (green) and deletions (blue). Structural variations (SV, size > 50 bp) in blue represent deletions, green bars are duplications (inversion). The purple line indicates variant novelty rate in percentage. Novelty is defined with respect to dbSNP141 for SNPs and indels whereas novelty for SV is determined by comparing to Database of Genomic Variants archive. **(B)** SNP and indel distribution by genomic features grouped by novelty. Variant distribution based on alternative allele frequency (common >5%, low \leq 5%) and variant effects predicted by SNPEff for SNPs **(C)** and indels **(D)**.

rates in genetic variant discovery, we determined that we could perform subsequent analyses using the identified genetic variants.

JIA novel variants discovery. We discovered a total of 1.49 million novel genetic variants. Specifically, we identified 10,800,221 autosomal bi-allelic single nucleotide polymorphisms (SNPs), of which 6,028,802 (55.88%) are common SNPs, defined as alternative allele frequency AAF >5%. However, 1,205,197 are novel or absent in dbSNP141 (11.2%, Fig. 1A, Table 1, Supplementary Table 2), and 2.14% of these novel SNPs (25,806) are common in all 48 JIA subjects. In addition, we identified 1,177,966 indels (novelty rate of 24%, 632,233 or 53.67% are common indels), 473,261 insertions (96,855 or 20.5% of which are novel) and 704,705 deletions (186,699 or 26.5% of which are novel, Supplementary Table 3). We then defined common structural variants (SVs) as those present in at least 16 out of 48 samples (i.e., present in 33% of the samples). From this analysis, 977 common SV (13.83% of the total novel SVs) were identified. Also, we found multiple novel SV (not reported in DGVA database) within

Impact	known		novel	
	AAF >5%	AAF ≤5%	AAF >5%	AAF ≤5%
SNPs	1,049	1,051	8	361
HIGH	1,003,906	627,315	2,329	199,137
LOW	22,431	24,837	204	5,377
MODIFIER	4,975,610	2,938,825	23,265	974,516
Indels				
HIGH	547	295	89	511
LOW	101,271	52,267	5,285	44,060
MODERATE	692	382	71	586
MODIFIER	495,305	243,653	28,973	203,979
	Singleton SNPs	Singleton indels	LoF SNPs	LoF indels
Average per sample	62,965	6,920	216	236
Maximum	120,411	13,196	233	274
Minimum	48,439	4,207	196	210

Table 1. Genetic variants discovered from whole genome sequencing on 48 JIA individuals. Number of JIA SNPs and indels classified by predicted functional impacts and alternative allele frequency (AAF). Lower panel summarizes per sample summary on singleton (private mutation) and loss of function (LoF) variants.

JIA subjects (Supplementary Figure 2). For example, we detected 5,545 deletions (DEL-678 novel), 398 inversions (INV-272 novel) and 1,120 duplications.

(DUP- 952 novel), resulting in an overall novelty rate of 26.9% (1,902/7,063) for SVs. When we stratified SV by type, we found novelty rates of 12.2%, 68.3% and 85% for DEL, INV and DUP respectively. High novelty rates for INV and DUP were expected, as the DGVA database contained relatively small proportions of INV (1.2%, 1,428) and DUP (17.2%, 21,360) from the total 120,045 SVs in its archive.

Next, we considered Loss of Function (LoF) variants, i.e., variants with potential to disrupt the function of protein coding genes. In total, we found 66,074 LoF variants (62,856 SNPs and 3,218 indels, Supplementary Figure 3, Supplementary Table 4). When we predicted the functional impact of LoF variants, we found 12,517 SNPs (1,924 novel) and 546 indels (320 novel) with potentially detrimental impacts. When we further examined the LoF variants with predicted detrimental effects to seek their regulation potentials, 932 SNPs and 2 indels were annotated with RegulomeDB class 1 regulatory evidence (eQTL, TF binding and DNase peaks, Supplementary Table 5).

We then examined the predicted regulation potential of SNPs discovered from WGS of JIA patients by mapping each SNP's genomic position to the RegulomeDB database. This procedure resulted in the assignment of SNPs to one of the 13 regulatory evidence classes (Supplementary Table 6). We observed significantly more JIA SNPs grouped under RegulomeDB class 1 (1a–f, the highest number of experimental findings supporting evidence of regulation) compared to 1KGP SNPs (Fisher T-test p-value $\leq 2.2E-16$ and 3.8 folds higher in JIA relative to 1KGP, Supplementary Figure 4). Despite the large sample size of 1KGP ($n = 2504$) versus our cohort ($n = 48$), we identified more SNPs with class 1 regulation evidence, corroborating the likelihood that these variants are part of the genomic regulatory architecture of the disease process.

In total we identified 40,754 common (alternative allele frequency, AAF $\geq 5\%$) and 10,504 low frequency (AAF $< 5\%$) JIA SNPs annotated with class 1 regulation evidence. To determine whether those SNPs were genuine mutations or sequencing artifacts, we compared the sequencing depth of 51,258 JIA SNPs with RegulomeDB class 1 regulation evidence (hereafter we denote these SNPs as class 1 JIA SNPs) and all other SNPs (that is, SNPs not annotated with class 1 evidence). This comparison showed that class 1 JIA SNPs exhibited a statistically significant higher sequencing depth than other SNPs (p-value $\leq 2.2e-16$). In addition, we examined variant quality and found higher quality scores in class 1 JIA SNPs compared with the other SNPs; higher quality scores indicate higher confidence calls (p-value $\leq 2.2e-16$). When we investigated the mapping quality of reads supporting class 1 JIA SNPs and other SNPs, we also identified statistical evidence for better mapping quality for class 1 JIA SNPs compared with the other SNPs (p-value $\leq 2.2e-16$). Taken together, the three metrics that quantified the fidelity of the detected SNPs (sequencing depth, variant quality and read mapping quality), make it reasonable to infer that class 1 JIA SNPs are genuine mutation in JIA genomes.

When we performed gene ontology (GO) enrichment analysis using genes of JIA SNPs with class 1 regulation evidence, significant GO terms included carboxylic acid metabolism and ATP hydrolysis coupled proton transport (Supplementary Table 7). These findings are provocative given the wealth of new data emerging regarding the importance of intermediary metabolism in regulating immune function⁷.

Genome wide distribution of JIA variants. WGS studies on the JIA samples demonstrated specific genetic variant hotspots within JIA genomes. Genes within these hotspots were annotated to be enriched in immune related biological processes. We defined hotspots as genomic regions (1 Mb bin size) that demonstrated a minimum of 1.5 fold greater density for JIA variants than for the 1KGP variants with Fisher exact test p-values less than 0.05. We identified 18 SNPs hotspots and 19 indel hotspots when comparing JIA WGS with 1KGP data (Fig. 2, Supplementary Table 8). We observed that those variant hotspots are significantly enriched relative to

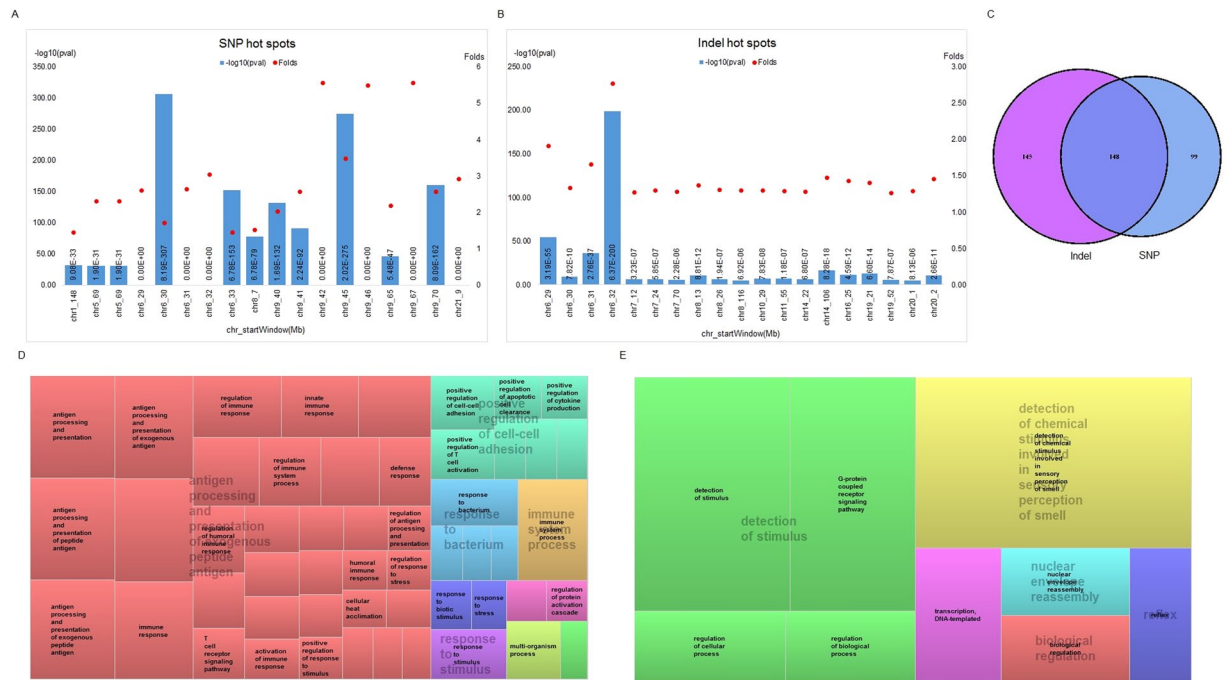


Figure 2. JIA variant hotspots per 1 Mb bin determined with respect to the 1000 Genomes Project variants. The bar shows Fisher exact test p values ($-\log_{10}$) of variant hotspots, and red dots show fold enrichment for each hotspot for (A) SNPs and (B) indels. (C) Venn diagram showing associated genes (overlapped or located within 5 kb upstream and downstream of hotspots) with SNP and indel hotspots. The 247 genes associated with SNP hotspots (of which 148 genes also associated with indel hotspots) were used for Gene Ontology (GO) enrichment analysis. The resulting GO term classifications are depicted using REVigo treemap in (D). GO enrichment analysis was also performed with 145 genes from indel hotspots. Results are shown in the GO term clustering represented by REVigo treemap in (E). Each rectangle of treemap indicates a GO terms cluster, sub-clusters (related GO terms) of same color are joined into super-clusters (centralized semi-transparent wordings). Size of rectangle reflects statistical significance p-values of GO terms, larger rectangles with smaller p-values.

1KGP with p-value range from $1.9E-31$ to 0. Genes associated with SNP hotspots were used as input for biological process GO enrichment analysis using GOrilla software⁸. We identified 46 significant GO biological process terms for genes associated with SNP hotspots (false discovery rate, $FDR \leq 0.05$, Supplementary Table 9). REVigo⁹ classified these GO terms into 8 groups. Predictably, we saw enrichment of immune-related biological processes (Fig. 2), compatible with the increasing evidence that suggests that the pathobiology of JIA involves complex interactions between innate and adaptive immunity¹⁰.

When we annotated genes associated with only indel hotspots, we found 8 significantly enriched GO terms for specific biological processes ($FDR \leq 0.05$, Supplementary Table 10). It is interesting to note that 34 genes associated with indel hotspots were enriched in G-protein coupled receptor signaling pathways (Fig. 2). This pathway involves in many pathophysiological processes including pathways related to arthritis¹¹, and we have recently demonstrated the importance of genes regulating G-protein coupled receptors in the transcriptional rewiring that characterizes treatment response in JIA¹².

We next characterized the profiles of SNPs within hotspots and found that novel SNPs have greater regulatory potential compared to known SNPs. In total, we identified 15,994 novel JIA SNPs within the 18 SNP hotspots (2,373 of which were found in at least in 33% of the cohort). Of these, 81 novel SNPs were annotated to have regulation potential (RP) scores ≥ 0.5 , significantly higher than known SNPs (p-value $5.96e-4$). When comparing SNPs within hotspots and SNPs outside hotspots, SNPs within hotspots have more regulation potential relative to SNPs outside hotspots (p-value $3.72e-27$). A representative genome browser view that depicts regulation properties of JIA variant hotspots is given in Supplementary Figure 5. This figure shows an overlapping of a hotspot region with histone marks (H3K27ac, H3K4me1, H3K27me3 and H3K4me3) found in three immune system related cell types (CD4+ T cells, CD20+ B cells and CD14+ monocytes). This same region has been identified as an expression quantitative trait locus (eQTL) for blood cells in the GTEx database¹³.

Association of JIA variants with regions of previously identified genetic risk in JIA. We next examined the new variants discovered on WGS in the context of previously identified risk loci for JIA. We found that JIA genetic variants, particularly those SNPs that co-localized within LD blocks containing previously identified JIA-associated SNPs, demonstrated higher regulation potential compared to SNPs outside the JIA risk-associated LD blocks.

The intersection between JIA variants with LD blocks containing regions of known genetic risk for JIA (hereafter denoted as LD blocks), as reported by Hinks *et al.*³ and Hersh *et al.*⁴ identified 9,423 SNPs (1,016 novel) and

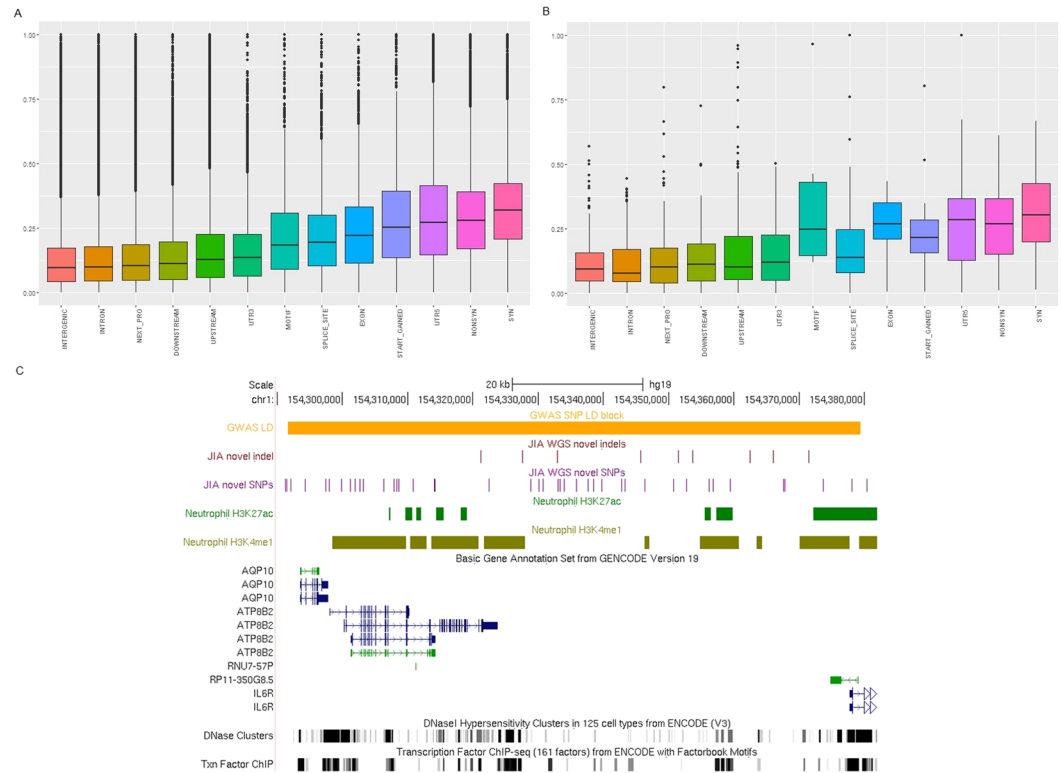


Figure 3. JIA SNPs within linkage disequilibrium (LD) blocks containing previously identified JIA-associated SNPs. Boxplot shows regulatory potential scores for SNPs overlapping blocks of JIA-associated SNPs grouped by genomic features for (A) all SNPs from WGS on 48 JIA individuals and (B) SNPs from WGS on 48 JIA individuals overlapping LD blocks containing previously identified JIA-associated SNPs. (C) Genome browser screenshot shows the LD block containing the SNP, rs11265608, intersecting novel JIA SNPs and novel JIA indels. Also, overlapping with neutrophil histone mark H3K27ac and H3K4me1. The gene annotation set from GENCODE v19 is shown by the rows below the H3K4me1 mark. The two rows of black vertical line at the bottom depict DNase hypersensitivity clusters in 125 cell types from ENCODE v3 and the transcription factor ChIP sequencing of 161 factors from ENCODE with factorbook motifs.

1,112 indels (261 novel, Supplementary Figure 6). We did not see evidence that the novel JIA variants were specifically enriched within these LD blocks. This was an unexpected result. Although, unlike candidate gene studies and the recently completed genetic fine mapping study³, the WGS approach is both unbiased and genome wide, we anticipated seeing new variants at a higher-than-expected frequency in those regions where genetic risk for JIA has already been established. As noted above, however, those SNPs that overlapped the JIA-associated LD blocks showed significant regulatory potential as assessed by the regulatory potential score database. We found significant enrichment of SNPs with RP scores ≥ 0.5 (p-value 0.0001 and 1.7 fold), indicating that SNPs co-segregating with JIA-associated LD blocks exhibited higher regulation ability compared to SNPs outside those LD blocks.

In order to determine whether the LD blocks were more immunologically active due to JIA genetic mutation or effects of variation from the general healthy cohort, we performed the same analysis on 1KGP data (healthy cohort). We overlapped 1KGP SNPs with LD blocks and obtained SNPs RP scores with reference to UCSC RP score databases. This was followed by generating a contingency table (for Fisher T-test analysis) that contained the number of 1KGP SNPs within (and outside) LD blocks, segregated by RP scores ≥ 0.5 and less than 0.5. We did not obtain statistical significant evidence of enrichment for the 1KGP SNPs within the LD blocks, i.e., having RP score ≥ 0.5 relative to those 1KGP SNPs outside LD blocks (p-value 0.6648). This finding supports the hypothesis that those LD blocks are more likely to be immunologically active due to the regulation potential of JIA SNPs, and are not generally impacted by SNPs from healthy subjects (1KGP).

In Fig. 3C, we present a representative genomic view of the region adjacent to the LD block containing the JIA-associated SNP, rs11265608³, which co-localized with novel SNPs and novel indels from WGS. We note that this region also overlaps with neutrophil H3K27ac and H3K4me1 signals⁶ and spans multiple transcription factor binding sites, suggesting that this is a particularly active genomic region.

In addition to querying the JIA-associated LD blocks, we examined genes (<https://ghr.nlm.nih.gov/condition/rheumatoid-arthritis#genes>) that are associated with rheumatoid arthritis (RA) and/or JIA (Supplementary Table 11, Supplementary Figure 7). We queried for enrichment of novel variants overlapping those associated genes. We did not see statistical significant of enrichment of novel SNPs/indels in arthritis associated gene regions.

Integration of JIA variants with epigenetic regulatory elements in CD4+ T cells, CD20+ B cells, and CD14+ monocytes. JIA is orchestrated by complex interactions between genetic and epigenetic mechanisms¹⁴. By leveraging publicly available epigenetic datasets from the ENCODE project¹⁵ and Roadmap Epigenomics project¹⁶, we compared JIA genetic variants to histone modification signals, open chromatin accessibility, DNase hypersensitivity sites and CTCF binding sites.

We started with histone modification integration analysis. We observed significant enrichment of JIA novel SNPs within H3K4me3 and H3K27me3 marks in CD20+ B cells (p-value 5.77e-114 and 2.63e-300 respectively). Both of these histone marks identify promoter activities¹⁷. We also observed that the novel JIA indels were enriched in H3K4me3 peaks of CD4+ T cells (p-value 1.98e-37). We did not observe enrichment of novel JIA variants in H3K27ac or H3K4me1 histone marks in any of the three cell types. In Supplementary Figure 8, we show a representative genome browser displaying novel JIA variants co-segregating with four histone signals (H3K27ac, H3K27me3, H3K4me1, H3K4me3) annotated in ENCODE and Roadmap Epigenomics data derived from CD20+ B cells, CD4+ T cells, and CD14+ monocytes at the chromosome 6 MHC region.

We next collated JIA SNPs/indels with DNase hypersensitivity sites in all three relevant cell types. DNase hypersensitivity sites mark open chromatin and thus are considered active regions with potential for regulating gene expression¹⁸. We found that novel JIA SNPs/indels were enriched in the DNase hypersensitivity sites of CD20+ B cells (p-value 3.98e121 and 4.88e-30 for SNPs and indels respectively). In addition, we found that novel JIA SNPs were enriched in CD14+ monocytes DNase hypersensitivity sites (p-value 2.09-e60). We did not see enrichment evidence of novel JIA variants within DNase hypersensitivity sites of CD4+ T cells. We provide an example of a genomic region where novel JIA SNPs/indels overlap DNase hypersensitivity sites of all three immune system cell types in Supplementary Figure 9. At this particular site, the nearest protein coding gene, LRRC37A (leucine rich repeat containing protein 37A), has the primary function of mediating protein-protein interactions and is involved in both innate immunity and nervous system development¹⁹. This gene is expressed in JIA neutrophils⁶ (maroon and light green rows of Supplementary Figure 9); note the presence of DNase hypersensitivity signals at adjacent loci (dark green peak signals).

Finally, we intersected JIA SNPs/indels with CTCF binding sites of three relevant cell types as a method for screening for potential underlying alterations in transcription factor binding sites due to genetic variation. This analysis yielded no statistical evidence of novel JIA variants enrichment within CTCF binding sites.

Discussion

The role of genetic risk for JIA has been well-established by family studies²⁰ as well as the higher than expected prevalence of other autoimmune or chronic inflammatory diseases in the families of affected children²¹. However, our knowledge of the genetic landscape for this family of illnesses remains incomplete. Furthermore, the two largest GWAS^{22,23} combined patients with 2 phenotypically different subtypes (oligoarthritis and polyarthritis), so subtype-specific data are still lacking.

Deep WGS provides an unprecedented opportunity to comprehensively study genetic landscapes at finer resolution than can be achieved with chip-based methods. While previous candidate gene approaches and genome-wide association studies have revealed useful information about genetic risk in JIA, a finer mapping is needed to decipher genetic landscapes of children with this disease to gain insights into pathogenesis. For example, even with dense genotyping using the Illumina Immunochip, Hinks *et al.*³ were able to explain no more than 18% of the risk for JIA. We therefore conducted deep WGS on 48 children with the polyarticular form of JIA over 2 independent cohorts, comparing results with publically available WGS data from healthy individuals (1KGP). The JIA specific genetic variation achieved an average of 72% concordance between the 2 cohorts despite the ethnic heterogeneity between them.

Using WGS, we were able to identify multiple new loci within JIA genomes that can now be considered candidate JIA risk regions suitable for further investigation. These loci consist of “hotspots” that were enriched for genetic variants discovered by WGS. Chromosome 6, where the major histocompatibility (MHC) genes are located, was a particularly prominent hotspot, supporting the existing view that JIA pathogenesis involves aberrant adaptive immune responses. However, the MHC locus is extraordinarily complex, and, in addition to genes that regulate the T cell responses, is rich in non-coding functional elements that include H3K4me1/H3K27ac-marked enhancers and non-coding RNA species in several relevant cell types²⁴. Thus, the presence of risk-associated variants within this genetically-rich region should not be taken as *prima facie* evidence that the causal variants include or are limited to MHC molecules. In the final analysis, functional studies in a broad spectrum of relevant cells will be required to identify causal variants. While this work has significantly expanded the variants that need to be queried, technologies now are available to accomplish this task²⁵.

This work corroborates the importance of the non-coding, functional genome in conferring risk for JIA. We have recently reported that the loci reported by Hinks *et al.*³ are enriched for H3K4me1/H3K27ac-marked enhancers in both CD4+ T cells and neutrophils⁶. In the current study, the newly-identified genetic variants were located within regions of functional elements in CD4+ T cells and CD20+ B cells as annotated by the ENCODE and Roadmap Epigenomics projects. Thus, JIA resembles almost every other complex trait, in which genetic risk lies largely in the functional, non-coding genome. This means that the field will be faced with the task of performing functional studies in a broad spectrum of cells, as there is no way to determine, a priori, the specific cell type in which genetic risk is most likely to operate. The appearance of JIA-associated genetic variants within functional elements in B cells, for example, was unexpected. Our findings invite additional inquiry into the role of B cells in this disease.

One finding of this JIA genomics study is that it corroborates the importance of previously identified JIA genetic risk loci. When we examined the newly identified genetic variants, those located within the LD blocks containing the known risk-associated SNPs had higher regulatory potential than variants situated outside those

LD blocks. On the other hand, when using 1KGP SNPs, we did not observe higher regulation potential within those LD blocks than 1KGP SNPs outside LD blocks, indicating that JIA SNPs within LD blocks very have functional significance. As noted above, however, our current work suggests that functional studies will need to be performed on a broad spectrum of cell types in order to clarify the causal link between genetic variance and immune dysfunction in JIA.

To date, this is the first deep WGS on JIA samples, serve as important data resource that has been instrumental in driving the progress biomedical and clinical research. We envision such data resource can often be benchmarked to deliver scientific impact in such a way that it facilitates the design of future hypothesis or experiment. Furthermore, genomic information from this data resource could lead to translational impact typically guides the changing of clinical practices, imposing clinical validity and industry relevance.

We are aware that these studies are just scratching the surface of where we want to go in building multi-dimensional genomic models for JIA. Such models will invariably include *disease-specific* epigenetic data, and concomitant genetic studies to determine whether and how underlying genetic variation alters epigenetic marks and genome function in pathologically-relevant cells. We are now learning that disease states are characterized by specific epigenetic signatures^{26, 27}. Recently, Peeters *et al.* identified²⁸ novel enhancer marks in memory T cells derived from synovial fluid of children with JIA. The degree to which these regulatory elements are essential to the disease process or even JIA-specific remains uncertain, however. Enhancer localization in inflammatory cells is highly dependent on the local microenvironment^{29, 30}, and it seems likely that similar findings would emerge from the study of inflammatory cells located at any site of chronic inflammation.

In summary, these studies demonstrate the utility of WGS, even on relatively small sample numbers, for elucidating underlying genetics of JIA. Segregation of variants from WGS of JIA individuals within known disease-associated LD blocks overcomes the limitation of depth and coverage of arrays in searching for JIA risk variants. Distribution of JIA SNPs/indels in regulatory elements serves as valuable resource from which to develop insights into epigenetic alterations underlying the effects of genetic variants. The long-term goal for this genomic study will be to integrate existing (and newly-developed) epigenetics resources and prospective JIA clinical records. Based on such data, “Precision Medicine” clinical decisions can be efficiently made to address clinical needs as well as answering important questions about JIA disease pathogenesis, clinical course, and the underlying biology of treatment response.

Materials and Methods

Samples. We performed WGS using Illumina X Ten on 50 samples from 48 individuals with JIA over two independent cohorts (Supplementary Table 12, 37 girls and 11 boys). All children fit International League Against Rheumatism criteria for rheumatoid factor negative, polyarticular JIA³¹. Fourteen of these children had detectable antinuclear antibodies at presentation. There were 29 samples in the first batch (B1), which included one pair of technical replicates (one subject was sequenced twice), while the second batch (B2) consisted of 19 subjects. All samples came from children of European or mixed European-American Indian ancestry and were obtained via a University of Oklahoma Institutional Review Board approved protocol. Informed consent was obtained from the parents of all patients. All research was carried out in accordance with the IRB-approved protocol and in compliance with relevant state and Federal regulations.

Sequencing data. Over all 48 samples, we obtained 39.81 billion properly paired-end reads (2×150 bp), aligned using Burrows-Wheeler Aligner (BWA)³² to the human reference genome GRCh37. An average of 97.87% reads per sample were mapped to the human reference genome (Supplementary Figure 10). A per sample average insert size of 352 bp was observed, meeting the targeted sequencing insert sizes of 300–400 bp (Supplementary Table 13). Furthermore, a mean sequencing depth at 38X across 48 samples, fulfilled the targeted 30X coverage (Supplementary Figure 11).

Variant discovery. We adopted the Genome Analysis Toolkit (GATK)^{33, 34} practices in calling SNPs and small insertions or deletions (indels, 1–50 bp). In order to identify high quality variants, we combined GATK variant quality score recalibration filtering and retained SNPs/indels that passed filtration criteria consisting of read depth $\geq 20X$, genotype quality ≥ 20 , variant quality ≥ 30 and minor-read ratio (MRR) ≥ 0.2 . To access variant novelty, we denoted novel variants (SNPs/indels) as those that were absent in dbSNP141 build 37³⁵ and known variants as those present in the dbSNP141.

For structural variation (SV, 100–100 Mbp), we used DELLY³⁶ to detect deletions (DEL), duplications (DUP), and inversions (INV), with reads of mapping quality ≥ 20 . We denoted novel structural variants (SVs) as those absent in Database of Genomic Variants archive (DGVa)³⁷.

WGS quality control. We first performed sample ethnicity screening by applying principal component analysis (PCA) over genotype information of JIA SNPs and 1KGP SNPs to compare our sample ethnicities against the reference to 1KGP data. This resulted in the JIA samples clustering together with the American samples of 1KGP (Supplementary Figure 12), as would be expected from the population distribution of the studied subjects.

We next examined per sample variant heterozygous to homozygous ratios (Het/Hom) to ensure that the ratio fell within the expected range of genomic scale at 1.5.³⁸ We computed an average of 1.6 Het/Hom per sample (Supplementary Table 14). We also examined transition to transversion ratios (Ti/Tv) for each sample, resulting in an average of 2.11 Ti/Tv per sample (Supplementary Table 15; the expected Ti/Tv is around 2.0³⁹).

Finally, for variant discovery QC, we examined the distribution of genetic variants across all autosomal chromosomes for each sample. We found that all samples demonstrate similar numbers of discovered variants (SNPs, indels and SV) at the same chromosomes (Supplementary Figure 13).

Variant annotation. SNPEff⁴⁰ was used for functional annotation of bi-allelic SNPs and indels. An in-house customized script was used to summarize the annotation results stratified by functional impacts (HIGH, MODERATE, LOW and MODIFIER), novelty, alternative allele frequency and genomic features (Supplementary Table 16).

In order to assess the potential functional significance of JIA SNPs, we mapped JIA SNPs to the regulatory potential (RP) scores⁴¹ downloaded from UCSC repository. The RP score measures similarity of patterns in alignments to those in known regulatory regions⁴². RP score ranges between 0 and 1, high score indicates high regulation potential.

Next, we annotated loss of function (LoF) variants following the definition used by MacArthur *et al.*⁴³. LoF SNPs were categorized to one of the following genomic features: non-synonymous, start lost, stop gained, stop lost and splice site. Indels that are classified as frameshift or splice sites are considered LoF indels. We predicted the functional effects of LoF SNPs using Polyphen⁴⁴ and categorized the prediction into two classes (damaging and benign). For LoF indels, we used the SIFT INDEL⁴⁵ bioinformatics tool to predict indels' functional effects as either damaging or benign. In addition, we conducted regulation potential predictions for LoF variants by mapping them to the RegulomeDB⁴⁶ database to obtain the associated regulation evidence (Supplementary Table 6).

Variant distribution. To investigate the distribution of SNPs and indels within JIA genomes, we computed variant density by counting number of variants in non-overlapping bins of 1 Mbp size over the entire genome. To identify variant hotspots, we compared JIA SNPs (indels) density and 1KGP SNPs (indels) density, i^{th} bin with the fold ratio given by:

$$\text{fold}_i = \frac{\text{num_JIA_Var_bin}_i / \text{total_JIA_var}}{\text{num_1KGP_Var_bin}_i / \text{total_1KGP_var}}$$

where num_JIA_Var_bin_{*i*} is the number of JIA variant in i^{th} bin and total_JIA_var represents genome wide total number of JIA variants. Similarly, num_1KGP_Var_bin_{*i*} is the number of 1KGP variant in i^{th} bin and total_1KGP_var represents genome wide total number of 1KGP variants, noting that 1KGP samples as healthy controls for this variant distribution analysis. We used a ratio of at least 1.5 and a Fisher exact test p-value ≤ 0.05 to define SNP/indel hotspots. For the detected variant hotspots, we searched for associated genes, arbitrarily designated as those genes located within the hotspot regions including genes located 5kbp upstream and downstream of hotspot regions. We then conducted gene ontology (GO) enrichment analysis using Gene Ontology enrichment analysis and visualization tool (GORilla)⁸ to assess the functions of the protein encoding genes within the variant hotspots. Reduce Visualize gene ontology (REVigo) was used to classify GO terms base on semantic similarity measurement⁹.

Variant association with regions of previously identified genetic risk for JIA. We examined JIA SNPs/indels that are located within regions of JIA genetic risk based on the previously identified LD blocks containing JIA-associated SNPs from studies by Hersh *et al.* and Hinks *et al.*^{3,4}. Bedtools⁴⁷ was used to intersect the discovered JIA SNPs/indels within those LD blocks (Supplementary Table 17). We then performed novel variant enrichment analysis using Fisher exact test. LD block information was obtained using the SNAP database⁴⁸ with reference to the 1KGP pilot1 at cutoff of $r^2 > 0.8$ and distance limit of 500 kilobases.

In addition, we collected a list of genes reported to have an association with arthritis from literature reviews (Supplementary Table 11) and intersected those genes with JIA SNPs/indels using bedtools.

Variant association with epigenetic regulatory elements of CD4+ T cells, CD20+ B cells and CD14+ monocytes. Interplays between epigenetic mechanisms and genetic variants may reveal important phenotype genotype relationships⁴⁹. Therefore, we chose pathologically relevant cell types, i.e., CD4+ T cells⁵⁰, CD20+ B cells⁵¹ and CD14+ monocytes⁵², to elucidate possible associations of JIA SNPs/indels and epigenetic elements (histone modifications, open chromatin and CTCF binding sites). We focused on four histone marks: H3K27ac, H3K4me1, H3K27me3 and H3K4me3, for integration with JIA genetic variants. For this analysis, we downloaded histone modification ChIP sequencing (ChIPSeq) data from GEO databases (Supplementary Table 18). Raw ChIPSeq reads were mapped to the human reference genome GRCh37 using BWA. Mapped reads were then used for peak calling by Model based Analysis of Chip-Seq (MACS2)⁵³, with default parameter settings for each histone mark independently. Next, we searched for co-localization between JIA SNPs/indels and chromatin accessible genomic regions. We obtained DNase hypersensitivity site data from the ENCODE project. We also investigated the co-segregation of JIA genetic variants and CTCF binding sites, which regulate gene expression through the organization of three dimensional chromatin structure⁵⁴. For CD4+ T cells we used CTCF binding site data identified by Cuddapah *et al.*⁵⁵ whereas for CD20+ B cells and CD14+ monocytes, we obtained the data from ENCODE (Supplementary Table 18).

We used bedtools to intersect the aforementioned epigenetic elements and JIA SNPs/indels. Enrichment analysis of novel SNPs/indels was conducted by Fisher exact test relative to known variants.

References

- Cobb, J. E., Hinks, A. & Thomson, W. The genetics of juvenile idiopathic arthritis: current understanding and future prospects. *Rheumatology* **53**, 592–599, doi:10.1093/rheumatology/ket314 (2014).
- Prahalad, S. Genetic analysis of juvenile rheumatoid arthritis: approaches to complex traits. *Current problems in pediatric and adolescent health care* **36**, 83–90, doi:10.1016/j.cppeds.2005.10.009 (2006).
- Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet* **45**, 664–+, doi:10.1038/ng.2614 (2013).
- Hersh, A. O. & Prahalad, S. Immunogenetics of juvenile idiopathic arthritis: A comprehensive review. *J Autoimmun* **64**, 113–124, doi:10.1016/j.jaut.2015.08.002 (2015).

5. Kitsios, G. D. & Zintzaras, E. Genome-wide association studies: hypothesis-“free” or “engaged”? *Translational research: the journal of laboratory and clinical medicine* **154**, 161–164, doi:10.1016/j.trsl.2009.07.001 (2009).
6. Jiang, K. Y. *et al.* Disease-Associated Single-Nucleotide Polymorphisms From Noncoding Regions in Juvenile Idiopathic Arthritis Are Located Within or Adjacent to Functional Genomic Elements of Human Neutrophils and CD4+ T Cells. *Arthritis Rheumatol* **67**, 1966–1977, doi:10.1002/art.39135 (2015).
7. Assmann, N. & Finlay, D. K. Metabolic regulation of immune responses: therapeutic opportunities. *The Journal of clinical investigation* **126**, 2031–2039, doi:10.1172/JCI83005 (2016).
8. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**, 48, doi:10.1186/1471-2105-10-48 (2009).
9. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS one* **6**, e21800, doi:10.1371/journal.pone.0021800 (2011).
10. Jiang, K. *et al.* Whole blood expression profiling from the TREAT trial: insights for the pathogenesis of polyarticular juvenile idiopathic arthritis. *Arthritis research & therapy* **18**, 157, doi:10.1186/s13075-016-1059-1 (2016).
11. Neumann, E., Khawaja, K. & Muller-Ladner, U. G protein-coupled receptors in rheumatology. *Nature reviews. Rheumatology* **10**, 429–436, doi:10.1038/nrrheum.2014.62 (2014).
12. Du, N. *et al.* Dynamic tracking of functional gene modules in treated juvenile idiopathic arthritis. *Genome medicine* **7**, 109, doi:10.1186/s13073-015-0227-2 (2015).
13. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and biobanking* **13**, 307–308, doi:10.1089/bio.2015.29031.hmm (2015).
14. Moncrieffe, H., Prahalad, S. & Thompson, S. D. Genetics of juvenile idiopathic arthritis: new tools bring new approaches. *Current opinion in rheumatology* **26**, 579–584, doi:10.1097/BOR.0000000000000094 (2014).
15. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, doi:10.1038/nature11247 (2012).
16. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330, doi:10.1038/nature14248 (2015).
17. Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell* **49**, 825–837, doi:10.1016/j.molcel.2013.01.038 (2013).
18. He, Y. *et al.* Genome-wide mapping of DNase I hypersensitive sites and association analysis with gene expression in MSB1 cells. *Frontiers in genetics* **5**, 308, doi:10.3389/fgene.2014.00308 (2014).
19. Giannuzzi, G. *et al.* Evolutionary dynamism of the primate LRR37 gene family. *Genome research* **23**, 46–59, doi:10.1101/gr.138842.112 (2013).
20. Prahalad, S. *et al.* Quantification of the familial contribution to juvenile idiopathic arthritis. *Arthritis and rheumatism* **62**, 2525–2529, doi:10.1002/art.27516 (2010).
21. Prahalad, S., Shear, E. S., Thompson, S. D., Giannini, E. H. & Glass, D. N. Increased prevalence of familial autoimmunity in simplex and multiplex families with juvenile rheumatoid arthritis. *Arthritis and rheumatism* **46**, 1851–1856, doi:10.1002/art.10370 (2002).
22. Thompson, S. D. *et al.* The susceptibility loci juvenile idiopathic arthritis shares with other autoimmune diseases extend to PTPN2, COG6, and ANGPT1. *Arthritis and rheumatism* **62**, 3265–3276, doi:10.1002/art.27688 (2010).
23. Thompson, S. D. *et al.* Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis and rheumatism* **64**, 2781–2791, doi:10.1002/art.34429 (2012).
24. Zhu, L. *et al.* Chromatin landscapes and genetic risk for juvenile idiopathic arthritis. *Arthritis research & therapy* **19**, 57, doi:10.1186/s13075-017-1260-x (2017).
25. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165**, 1519–1529, doi:10.1016/j.cell.2016.04.027 (2016).
26. Seumois, G. *et al.* Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nat Immunol* **15**, 777–788, doi:10.1038/ni.2937 (2014).
27. Zhu, H. *et al.* Whole-genome transcription and DNA methylation analysis of peripheral blood mononuclear cells identified aberrant gene regulation pathways in systemic lupus erythematosus. *Arthritis research & therapy* **18**, 162, doi:10.1186/s13075-016-1050-x (2016).
28. Peeters, J. G. *et al.* Inhibition of Super-Enhancer Activity in Autoinflammatory Site-Derived T Cells Reduces Disease-Associated Gene Expression. *Cell Rep* **12**, 1986–1996, doi:10.1016/j.celrep.2015.08.046 (2015).
29. Gosselin, D. *et al.* Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell* **159**, 1327–1340, doi:10.1016/j.cell.2014.11.023 (2014).
30. Lavin, Y. *et al.* Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326, doi:10.1016/j.cell.2014.11.018 (2014).
31. Petty, R. E. *et al.* International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J Rheumatol* **31**, 390–392 (2004).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, doi:10.1093/bioinformatics/btp324 (2009).
33. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303, doi:10.1101/gr.107524.110 (2010).
34. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491–+, doi:10.1038/ng.806 (2011).
35. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311, doi:10.1093/Nar/29.1.308 (2001).
36. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339, doi:10.1093/bioinformatics/bts378 (2012).
37. Lappalainen, I. *et al.* DbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res* **41**, D936–941, doi:10.1093/nar/gks1213 (2013).
38. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947, doi:10.1038/nature08795 (2010).
39. Liu, Q. *et al.* Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC genomics* **13**(Suppl 8), S8, doi:10.1186/1471-2164-13-S8-S8 (2012).
40. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92, doi:10.4161/fly.19695 (2012).
41. Kolbe, D. *et al.* Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome research* **14**, 700–707, doi:10.1101/gr.1976004 (2004).
42. Taylor, J. *et al.* ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. *Genome research* **16**, 1596–1604, doi:10.1101/gr.4537706 (2006).
43. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828, doi:10.1126/science.1215040 (2012).
44. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–249, doi:10.1038/nmeth0410-248 (2010).

45. Hu, J. & Ng, P. C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS one* **8**, e77940, doi:10.1371/journal.pone.0077940 (2013).
46. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome research* **22**, 1790–1797, doi:10.1101/gr.137323.112 (2012).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, doi:10.1093/bioinformatics/btq033 (2010).
48. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939, doi:10.1093/bioinformatics/btn564 (2008).
49. Walsh, A. M. *et al.* Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. *Genome Biol* **17**, doi:ARTN 7910.1186/s13059-016-0948-6 (2016).
50. Ito, Y. *et al.* Detection of T cell responses to a ubiquitous cellular protein in autoimmune disease. *Science* **346**, 363–368, doi:10.1126/science.1259077 (2014).
51. Bugatti, S., Vitolo, B., Caporali, R., Montecucco, C. & Manzo, A. B cells in rheumatoid arthritis: from pathogenic players to disease biomarkers. *BioMed research international* **2014**, 681678, doi:10.1155/2014/681678 (2014).
52. Davignon, J. L. *et al.* Targeting monocytes/macrophages in the treatment of rheumatoid arthritis. *Rheumatology* **52**, 590–598, doi:10.1093/rheumatology/kes304 (2013).
53. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
54. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions (vol 453, pg 948, 2008). *Nature* **500**, 242–242, doi:10.1038/nature12316 (2013).
55. Cuddapah, S. *et al.* Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research* **19**, 24–32, doi:10.1101/gr.082800.108 (2009).

Acknowledgements

We thank all the subjects and their families in this study for their participation. This work was supported by R01-AR-060604 from the National Institutes of Health and the USA Innovate Research Grant from the Arthritis Foundation.

Author Contributions

Laiping Wong performed computational analysis of the DNA WGS data and wrote the manuscript. Kaiyu Jiang assisted in sample preparation and assisted the writing of the manuscript. Yanmin Chen assisted in samples preparation. James Jarvis designed the study and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-02966-9

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017