medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Whole Genome Sequencing identifies novel structural variant in a large Indian family affected with X - linked agammaglobulinemia

Abhinav Jain[1,2,#], Geeta Madathil Govindaraj[3,5,#,$], Athulya Edavazhippurath[3,4], Nabeel Faisal[3], Rahul C Bhoyar[1], Vishu Gupta[1,2], Ramya Uppuluri[6], Shiny Padinjare Manakkad[3], Atul Kashyap[1], Anoop Kumar[1], Mohit Kumar Divakar[1,2], Mohamed Imran[1,2], Sneha Sawant[7], Aparna Dalvi[7], Krishnan Chakyar[3], Manisha Madkaikar[7], Revathi Raj[6], Sridhar Sivasubbu[1,2,$], Vinod Scaria[1,2,$]

[1]CSIR-Institute of Genomics and Integrative Biology, Mathura Road, New Delhi, Delhi, India
[2]Academy of Scientific and Innovative Research (AcSIR), Mathura Road, New Delhi, Delhi, India
[3]Department of Pediatrics, Government Medical College Kozhikode, Kozhikode, Kerala, India
[4]Multidisciplinary Research Unit, Government College Kozhikode, Kozhikode, Kerala, India
[5]FPID Regional Diagnostic Centre, Government Medical College Kozhikode, Kozhikode, Kerala India
[6]Department of Pediatric Hematology, Oncology, Blood and Marrow Transplantation, Apollo Hospitals, 320, Padma complex, Anna salai, Teynampet, Chennai, Tamil Nadu, India
[7]Department of Pediatric Immunology and Leukocyte Biology, ICMR-National Institute of Immunohaematology, KEM Hospital, Parel, Mumbai, Maharashtra, India.

 # Joint first authors

$ Corresponding author

Vinod Scaria: vinods@igib.in

Sridhar Sivasubbu: sridhar@igib.in

Geeta Madathil Govindaraj: geetakkumar@gmail.com

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Abstract

Background

X - linked agammaglobulinemia (XLA, OMIM #300755) is a primary immunodeficiency disorder caused by pathogenic variations in the *BTK* gene, characterized by failure of development and maturation of B lymphocytes. The estimated prevalence worldwide is 1 in 190,000 male births. Recently, genome sequencing has been widely used in difficult to diagnose and familial cases. We report a large Indian family suffering from XLA with five affected individuals.

Methods

We performed complete blood count, immunoglobulin assay, and lymphocyte subset analysis for all patients and analyzed Btk expression for one patient and his mother. Whole exome sequencing (WES) for four patients, and whole genome sequencing (WGS) for two patients have been performed. Carrier screening was done for 17 family members using Multiplex Ligation-dependent Probe Amplification (MLPA) and haplotype ancestry mapping using fineSTRUCTURE was performed.

Results

All patients had hypogammaglobulinemia and low CD19+ B cells. One patient who underwent Btk estimation had low expression and his mother showed a mosaic pattern. On structural variant analysis of WGS data, we found a novel large deletion of 5,296 bp at loci chrX:100,624,323-100,629,619 encompassing exons 3-5 of the *BTK* gene. Family screening revealed seven carriers for the deletion. Two patients had a successful HSCT. Haplotype mapping revealed mainly South Asian ancestry.

Conclusion

Whole genome sequencing led to identification of the accurate genetic mutation which could help in early diagnosis leading to improved outcomes, prevention of permanent organ damage and improved quality of life, as well as enabling prenatal diagnosis.

Keywords: X linked Agammaglobulinemia, Next Generation Sequencing, Structural Variant, Family Screening, Multiplex Ligation-dependent Probe Amplification, Hematopoietic Stem Cell Transplantation, South Asian Ancestry,

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Introduction

X - linked agammaglobulinemia (XLA) is a monogenic primary immunodeficiency disorder caused by pathogenic mutations in the *BTK* (Bruton's tyrosine kinase) gene [1] with X-linked recessive inheritance. The *BTK* gene is involved in the development, maturation, and signaling of B cells [2]**.** The absence of plasma B cells leads to markedly reduced levels of all classes of immunoglobulins .Patients with pathogenic variants in the *BTK* gene typically manifest with recurrent infections between 3 and 18 months of age. The commonest infections are those of the respiratory tract caused by encapsulated bacteria. [3]. The [4] prevalence of XLA has been estimated to be 1 in 190,000 male births or 1 in 379,000 total live births, with 40% of cases having a positive family history [5]. As per the diagnostic criteria, the patient needs to be a male, who has hypogammaglobulinemia or agammaglobulinemia , <2 percent CD19+ B cells, and either a male family member of maternal lineage who is documented to have agammaglobulinemia and <2 percent CD19+ B cells or a confirmed (by DNA, messenger ribonucleic acid (mRNA), or protein analysis) defect in the *BTK* gene or Btk expression [6].

The BTK gene loci are mapped to Xq21.3-Xq22 spanning 36.7 kb on the X chromosome and consists of 19 exons and 5 functional domains. A large number of genetic variants spanning this locus have been mapped to the *BTK* loci and systematically collected and deposited in the BTKbase database http://structure.bmc.lu.se/idbase/BTKbase/ [7]. A large number of mutations reported are single nucleotide variations (73%) that include missense, nonsense, and splice site variations. Small insertions and deletions (18%) and structural variations (9.5%) account for the remaining mutations. Only 3.5% of the gross deletions reported disrupting the functionality of the *BTK* gene [8]. A large number of patients with XLA have been reported with either a deletion in only the *BTK* gene or gross deletions encompassing *BTK* along with a few neighboring genes like *TIMM8A, TAF7L, Artemis, IGHM,* and *DRP2 [9, 10].* The contiguous deletion in the *BTK* and *TIMM8A* gene that causes immunodeficiency with dystonia, optic neuronopathy, and sensorineural deafness is known as XLA-Mohr-Tranebjærg syndrome (XLA-MTS) [11]

We report five patients belonging to a large Indian family suffering from XLA. Whole genome sequencing identified a novel large deletion of 5,296 bp at loci chrX:100,624,323-100,629,619 spanning *BTK* exons 3-5. Screening by Multiplex

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Ligation-dependent Probe Amplification (MLPA) in the extended family identified seven carriers for the deletion. We surmise that the approach from next-generation sequencing to a low-cost screening method is replicable and could help to reduce the disease burden in the community.

MATERIALS AND METHODS

Patient and Clinical Workup

Five male children belonging to an Indian family were evaluated as a part of a programme on primary immune deficiency disorders at the Government Medical College, Kozhikode, Kerala, and the Institute of Genomics and Integrative Biology, Delhi between 2015 and 2019. The study was approved by the Government Medical College, Kozhikode by the Institutional Ethics Committee (Ethics No. GMCKKD/RP2017/IEC/147). The index cases were P2 and P3. After elaborate pedigree analysis, clinical characteristics were recorded in a semi-structured proforma and included age at diagnosis, number of hospitalizations and PICU admissions, type of infections, complications, outcome, etc. Clinical investigations performed were complete blood count, lymphocyte subset analysis, and immunoglobulin assay.

Btk Protein Expression

The cytoplasmic staining procedure was used to analyze the Btk protein expression by flow cytometry. We took 50 µL of whole blood from the patient (P2), mother, and a healthy unrelated control which was surface stained using CD3-FITC, CD14-PE for 30 min at 37^0C. An unstained tube was maintained for all three samples. At the end of incubation, these were fixed with formaldehyde for 10 minutes followed by permeabilization with Triton-X for 30 minutes at 37^0C. Subsequently, these were washed, stained with an anti-Btk monoclonal antibody for 45 minutes, washed, and analyzed by flow cytometry.

DNA Isolation and Whole Exome Sequencing

Approximately 5 ml of blood from patients (P1, P2, P4, and P5) and their family members were drawn in an acid citrate dextrose (ACD) tube (Becton Dickinson, NJ, USA). Due to the unexpected death of P3, we could not collect his sample. Genomic DNA was isolated using the salting-out method [12] and 100 ng was used as a template to perform WES on the patient samples using the Truseq Exome library prep kit following the standard procedure

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

as per the manufacturer's protocol (Cat no.: 20000408, Illumina Inc., SA USA). The prepared library ~~was~~ sequenced on the HiSeq2500 for three patients (P1, P2, and P4) and the NovaSeq6000 platform for P5 patients had paired-end read with a read length of 150 bp.

Exome Data Analysis

The raw reads underwent quality control at Phred score Q30 using Trimmomatic-0.38 [13]. The trimmed reads were mapped on to human reference genome GRCh37 using Burrows-Wheeler Aligner (BWA) version 0.7.17 [14]. The mapped reads were further sorted and duplicate reads were removed using SAMtools [15] and Picard respectively https://broadinstitute.github.io/picard/. The variant calling was done using the HaplotypeCaller of Genome Analysis ToolKit (GATK version 3.8.0) best practices [16]. The variants were further systematically annotated using ANNOVAR (2018-04-16 00:47:49) [17] that comprises of multiple datasets i.e. refGene, dbsnp (avsnp150), dbnsfp35a, and clinvar 20190305**.** It also annotates allele frequency from the global population datasets i.e. 1000 Genome project (1000g2015_all), Genome Aggregation Database (gnomAD V2.1.1), and Esp6500. The protein altered variants (missense, stop gain, stop loss, frameshift, non-frameshift, splicing, small indels, small insertion, and small deletion) were prioritized. Further, variants whose minor allele frequency was greater than 5% in the global population were filtered out. A gene filter was applied which comprised 454 genes known to be implicated in PID and were recently catalogued by the International Union of Immunological Societies (IUIS) expert committee [18]. Finally, based on the phenotype, we correlated the variant's clinical significance.

Whole Genome Sequencing and Analysis

The whole genome sequencing was performed using 100 ng of genomic DNA from two patients (P1 and P4). WGS was performed with 150 bp paired-end reads that were generated using Truseq PCR free library kit as per manufacturer's instructions (Cat no.: FC-121-9006DOC Illumina Inc. SA USA) on Illumina NovaSeq 6000 platform (San Diego, CA, USA) using sequencing by synthesis chemistry. The raw reads to variant annotations were performed similarly to the whole exome analysis. We merged both the individuals' variants using the GATK (version 3.8.0) tool called CombineVariants. We filtered out the variants with MAF>0.05 in the global population. Further, we adopted an overlap based strategy [19]

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

where we prioritized the common homozygous variants between P1 and P4. Further, variants were filtered based on the in-silico tool CADD (Combined Annotation-Dependent Depletion) score > 15 [20]. For the remaining variants, we manually correlated phenotypic characteristics of patients with filtered variants.

We have also performed structural variant (SV) analysis on the whole genome sequenced aligned reads of both the patients (P1 and P4). We used LUMPY (version 0.2.13) for SV calling [21] that were further genotyped using SVtyper [22]. The SVs were further prioritized using an overlap-based strategy where we prioritized a common homozygous variant in the 454 PID genes [19]. To validate the SV result, we adopted the manual coverage-based analysis on WGS paired-end reads that were aligned on the reference genome hg19/GRCh37 using Integrated Genome Viewer (IGV) [23].

Multiplex-ligation dependent probe amplification (MLPA) assay

To identify and validate the gross deletion in extended family members, the MLPA based approach was adopted. We have performed the test on 17 members of the family with their consent and institutional ethical approval (Ethics No. GMCKKD/RP2017/IEC/147). Genomic DNA was isolated using a standard salting-out method [12] and 100 ng of the genomic DNA was used. MLPA was performed as per manufacturer's instructions (MRC-Holland, Amsterdam, The Netherlands) using SALSA MLPA EK1 reagent kit-FAM (EK1-FAM, MRC Holland, Netherlands) along with the SALSA MLPA Probemix P210 BTK (P210-050R, MRC Holland, Netherlands). Capillary electrophoresis of the amplicons was performed on ABI 3130 genetic analyser (Applied Biosystems™, California, USA). MLPA data analysis was performed using Coffalyser.net software (MRC Holland, Netherlands).

Ancestry Haplotype Mapping using Chromosomal Painting

Since it is a novel deletion, we explored the haplotype similarity pattern of our patients with that of the global population. We used the haplotype prediction tool fineSTRUCTURE [24] (version 2.1.3) on P1 and P4 whole genome sequenced variants, whose chrX variants have been extracted and merged it with chrX variants of 1000 Genome Project as a reference and 44 whole genome data of the Qatar population. The 1000 Genome project consists of 2504 individuals from five major populations African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SOU) [25]. We pruned the merged VCF by applying

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

a variant filter of allele frequency greater than 1% and allele number filter of 1,000 to get the maximum genotype rate using a bespoke bash script. We phased the merged VCF using the SHAPEIT v2.r900 tool [26]. We ran fineSTRUCTURE pipeline which involves four steps that include painting of chromosome depending on the population haplotype with the individual i.e. chromopainter then combining all painted data and assigning a population to a block which involves chromocombine and fineSTRUCTURE respectively, and finally tree building based on maximum posterior population inference. The region 50KB, 500KB, 5MB, and 50MB upstream and downstream on both sides of the chromosomal locus chrX:100624323-100629619 has been plotted (hg19/GRCh37) using R scripts provided by fineSTRUCTURE.

RESULTS

Clinical details

The five male children presented to the Government Medical College, Kozhikode, Kerala with recurrent respiratory infections, diarrhea, pyoderma, and pyogenic meningitis with onset between 5 months and 12 months of age. They belonged to a large extended family. Clinical features have been tabulated in **Table 1**.

All the children had absent or atrophic tonsils, hypogammaglobulinemia, severely reduced or absent CD19 counts, and were started on intravenous immunoglobulin (IVIG) infusions. Two children who presented late, and were not on regular prophylaxis with IVIG developed bronchiectasis. Other complications included stunted growth, delayed puberty, and arthritis. One child died soon after diagnosis due to meningitis at 20 months of age. Analysis of the pedigree chart revealed that 22 male children had died in the early years of life.

Btk expression estimation using flow cytometry

By using this flow cytometric approach, patient P2 has low Btk expression observed on monocytes (37%) compared to 87% in the control, the normal range being 90 +/- 5%. His mother showed a mosaic pattern suggestive of being a carrier for XLA.

Whole exome sequencing analysis

We performed whole exome sequencing for four patients (P1, P2, P4, and P5) with 99.5% alignment on the human reference genome for each patient and average coverage of

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

90.8X. A total of 467,638 variants were called for each patient. Variants were annotated using a tool ANNOVAR [17] on average, 13,074 protein-altering variants consisting of splicing, exonic splicing, and exonic variants except synonymous SNVs were prioritized. These variants were further prioritized for minor allele frequency (MAF) < 5% in the global population which reduces the average number of variants to 2,383. On average 12 genetic variants were mapped to 454 primary immunodeficiency genes, for further downstream analysis. We could not find any variant which could correlate with the clinical features. The whole exome sequencing data with variant filtering has been tabulated in **Supplementary Table 1**.

Whole Genome Sequencing Analysis

Since we could not find any causal variant using whole exome sequencing, we performed whole genome sequencing of P1 and P4 with a mapping percentage of 99.58% and 99.49% and coverage of 21X and 86X respectively. Variant calling was done by GATK (version 3.8.0) HaplotypeCaller and resulted in 4,905,687 and 4,904,616 variants for P1 and P4 respectively. The variant files of both the patients P1 and P4 were merged and annotated using a tool ANNOVAR [17] which led to a total of 6.5 million variants. These variants were prioritized for minor allele frequency (MAF) < 5% in the global population dataset, that reduces the variant number to 890,387. Further, we adopted an overlap-based strategy [19] and prioritized 27,236 variants. On applying in-silico computational tool i.e. CADD score > 15, we prioritized 4 variants. On manual correlation with the filtered variants, we could not correlate any of the prioritized variants with the clinical characteristics of the patients. The whole genome sequencing data with variant filtering has been tabulated in **Supplementary Table 2.**

We have also performed structural variant analysis using LUMPY for calling SV. We got a total of 8,925 and 29,638 SVs in patient P1 and P4 respectively. Prioritizing common and homozygous SVs in both the patients, led to a total of 768 SVs. Further we applied the PID gene filter that drops the number to 8 SVs. Finally, on clinical correlation with SVs, we prioritized a novel large deletion of 5,296 bp on chrX at loci ranging from 100,624,323 to 100,629,618. The variant filtering at each step has been tabulated in **Table 2.** Further visualizing on the human reference genome hg19/GRCh37 using IGV, we found that there were no reads present on chrX at loci ranging from 100,624,323 to 100,629,618 i.e. 5296

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

bases (~5 Kb). This led to the identification of a large deletion spanning exon 3 to exon 5 of the *BTK* gene for both P1 and P4 as shown in **Figure 1**. However, the region flanking the deleted loci showed adequate mapped reads. The deleted region in the patient has also been properly covered in the control sample. This prompted us to visualize the whole exome data for P2 and P5 on IGV and as a result, we found the same *BTK* gene deletion encompassing exon 3-5 as shown in **Supplementary Figure 1**. On intersecting patients deleted loci chrX:100624323-100629619 with the 1000 Genome project SVs, gnomAD SVs, and IndiGen SVs (in-house) database, we could not find any structural variant that falls in the exon 3-5 of *the BTK* gene in any of the databases.

Hematopoietic Stem Cell Transplantation (HSCT)

After the family was counseled regarding the feasibility of HSCT as a curative option, two patients P2 and P4 underwent HSCT from a matched sibling donor. They required myeloablative conditioning to prevent graft rejection, and a treosulfan-based reduced toxicity protocol to ensure adequate myeloablation. Graft versus host disease prophylaxis consisted of short-course methotrexate and tacrolimus. Both of them are now disease-free.

Family screening using the multiplex ligation-dependent probe amplification (MLPA) assay

We performed whole genome sequencing on two samples P1 and P4 and found a large hemizygous deletion encompassing 3-5 exons of the *BTK* gene. To confirm whether the mutation is *de novo* or inherited, MLPA assay-based approach was adopted. The assay was first validated using the P1 and P4 samples as positive controls and two reference samples from healthy individuals. The positive control samples P1 and P4 were found to have the probe ratio of zero for the exons 3-5 of *BTK* gene i.e. there is the absence of any copy for this region as shown in **Figure 2** and therefore, the deletion was detected, which corroborated with the whole genome sequencing results shown in **Figure 1**.

The standardized and validated MLPA assay was then used to screen additional family members of P1 and P4 for the deletion in the *BTK* gene. The probe ratios for the mothers of both P1 and P4 corresponded to ~0.5 as shown in **Supplementary Figure 2**, therefore, they were heterozygous for the same deletion. This shows that the deletion has been inherited by both P1 and P4 from their parents. The sibling and parent of P4 have also been tested

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

and were found to have a normal copy of the gene. The other two patients, P2 and P5 were tested using the MLPA based assay and were found to harbor the same deletion in hemizygous form as shown in **Figure 2**. Following this, the families of P2 and P5 were tested. The maternal grandparent, parent, and the two siblings of P2 were found to be heterozygous for the deletion, and one sibling was found to be carrying a normal copy of the gene **Supplementary Figure 2**. The parent of P5 was found to be heterozygous for the deletion and her two siblings were found to have inherited both the normal copies of the gene. Additionally, two other members of the extended family were tested for the deletion and were found to have normal copies of the gene. Using MLPA-based assay, it has been found that the large deletion is inherited in multiple members of the family. Out of 17 members tested, we found four members of the family were hemizygous and seven were carriers for the deletion. Upon analysis of the pedigree, we can predict more than half a dozen of the family members could be carriers of the deletion. The MLPA probe ratio for each individual at 19 *BTK* exons test probes as well as for reference probes has been tabulated in **Supplementary Table 3.**

Chromosomal painting analysis

Since this variation is absent in all the control populations, it excites us to know the haplotype ancestry around the deleted region. In order to predict haplotype ancestry of the region flanking the deletion on chromosome X, we merged chrX variants of whole genome sequenced P1 and P4 individually with 2,504 individuals of five major populations from 1000 Genome Project and 44 whole genome sequenced Qatari individuals which comprise 3,821,263 variants. We pruned the variants whose allele frequency is less than 0.01 and allele number less than 1,000 that filtered the variant number to 185,258, and 186,417 for P1 and P4 respectively. After phasing and chromosomal painting using fineSTRUCTURE, we found that both the samples had South Asian ancestry. The painted chromosomal region of both the patients P1 and P4, 5MB upstream and downstream of the deleted locus has been well represented in **Figure 3.** For more fine visualization we have painted the chromosomal region at 50KB, 500KB, and 50MB upstream and downstream to the deleted locus in **Supplementary Figure 3.**

Discussion

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

XLA was the first primary immunodeficiency disorder discovered by a pediatrician Ogden Carr Bruton in 1952 [27]. The onset of symptoms is usually before one year of age. Infections do not occur in early infancy due to the protective effect of maternally derived IgG. [5]**.** Apart from an increased frequency and severity of respiratory and gastrointestinal infections, there are other rare manifestations described including arthritis, which was a feature in P5, and significantly impinges on the quality of life [28]. Autoimmunity and autoinflammation can occur in XLA, and arthritis could be either due to a dysfunctional immune system or due to infection with mycoplasma, which was not the case in our patient P5, [28]**.** The challenge lies in identifying the index case[29]. Although panhypogammaglobulinemia is the rule, there are reports of normal immunoglobulin levels or selective immunoglobulin deficiency [30, 31]. Antibody responses to isohemagglutinins and vaccine antigens were not performed since B cell numbers were negligible, and it was possible to look for a *BTK* variant.

Identifying the variant is critical for carrier detection and genetic counseling, without which the disease frequency in the community cannot be reduced. We have performed WES in all the four patients but unfortunately were unable to find the causal variant. This could be due to the limitation of WES in providing adequate coverage for all the genes [32] inability in calling structural variants [33]**,** missing non-coding variants [34] as well as high-quality coding SNVs, [35]**.** The WGS has paved the way for the diagnosis of such difficult-to-diagnose disorders. In the present study, we identified a novel large deletion of 5,296 bp encompassing exons 3-5 of the *BTK* gene in four patients. A very recent analysis was done by Thaventhiran and group, where they performed WGS analysis on 1,318 PID patients and identified eight structural variants which could have been missed by WES [36]. WGS also inferred *denovo* 3Mb deletion breakpoints in two probands affected by DiGeorge syndrome [37].

XLA is caused by pathogenic variants in the *BTK* gene, which is a cytosolic tyrosine kinase protein composed of five domains i.e. pleckstrin homology (PH), Tec homology (TH), Src homology domain (SH2), SH3, and kinase (Catalytic) domains [38]. In our patients, the 5,296 bp deletion spans exons 3-5 of *BTK* gene. Exon 3 is an integral part of the PH domain, which regulates the binding of molecules like inositol 1,3,4,5-tetrakisphosphate (IP_4), inositol 1,3,4,5,6-pentakisphosphate (IP_5), and inositol 1,2,3,4,5,6-hexakisphosphate (IP_6) which in

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

turn activates *BTK* gene function [39]**.** There are multiple mutations reported in exon 3 which affect the functionality of the PH domain [40] and cause XLA. Mutations in exon 4 which encode for the C-terminal immunoglobulin domain [41] and mutations in exon 5 also have been reported in a patient with decrease in Btk expression and CD19+ B cell number [42].

Early diagnosis is imperative to avoid morbidity, mortality and vaccination with OPV, since these children are at risk for paralytic poliomyelitis. [43]. In the family studied, 22 male children had died due to infections. These individuals might have been harbouring the same hemizygous deletion and early diagnosis could have been life-saving. . Using a low-cost MLPA-based assay for screening additional family members, we have found seven female carriers for the *BTK* exon 3-5 deletion. Unfortunately, we were able to test only 17 out of 159 individuals. But by analyzing the pedigree, we can predict that more than half a dozen family members could be carriers. Carrier detection in families affected by Mendelian disorders will enable genetic counseling and antenatal diagnosis, ultimately resulting in a reduced disease burden. [44, 45]. MLPA has been used to screen for multiple diseases as well as prenatal screening, due to its swift, highly sensitive and cost-effective approach [46, 47]. Both children in whom the diagnosis was delayed and who were not on regular IVIG prophylaxis (P1 and P5) developed permanent lung damage.

There are reports of successful HSCT for XLA, but it has not been used extensively since immunoglobulin replacement is widely available. However, when the cost and availability of lifelong prophylaxis is a limiting factor, HSCT has been chosen by parents instead of the option of lifelong immunoglobulin replacement [48, 49]. This was the case in P2 and P4 patients.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

While the mean age at onset was 7 months, the mean age at diagnosis was 44.6 months, a delay of 37.6 months, resulting in multiple hospital admissions for treatment of infections. Clinician and patient/parent education would help to ensure early diagnosis, enhance compliance with treatment and prevent poor outcomes as occurred in P1 and P5 . Transitioning to adult services is also a challenge [50]. Accurate genetic workup and counseling of the extended family will reduce the burden of care. [51].

The WGS and the low-cost MLPA assay are replicable in the society for reducing the disease burden of affected families and the community at large. The effectiveness of this approach hinges on the availability and accessibility of a system for genetic counseling that the community would accept.

References

1 Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S. Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. *Cell* 1993;**72**:279–90.

2 Maas A, Hendriks RW. Role of Bruton's tyrosine kinase in B cell development. *Dev Immunol* 2001;**8**:171–81.

3 Moin M, Aghamohammadi A, Farhoudi A, Pourpak Z, Rezaei N, Movahedi M, Gharagozlou M, Ghazi BMS, Zahed A, Abolmaali K, Mahmoudi M, Emami L, Bashashati M. X-linked agammaglobulinemia: a survey of 33 Iranian patients. *Immunol Invest* 2004;**33**:81–93.

4 Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S. Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. *Cell* 1993;**72**:279–90.

5 Winkelstein JA, Marino MC, Lederman HM, Jones SM, Sullivan K, Burks AW, Conley ME, Cunningham-Rundles C, Ochs HD. X-linked agammaglobulinemia: report on a United States registry of 201 patients. *Medicine* 2006;**85**:193–202.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

6 Zwebb, www. zwebb.com. ESID - European Society for Immunodeficiencies. https://esid.org/Working-Parties/Clinical-Working-Party/Resources/Diagnostic-criteria-for-PID2#Q15 (accessed 12 Sep2020).

7 Väliaho J, Smith CIE, Vihinen M. BTKbase: the mutation database for X-linked agammaglobulinemia. *Hum Mutat* 2006;**27**:1209–17.

8 Conley ME, Broides A, Hernandez-Trujillo V, Howard V, Kanegane H, Miyawaki T, Shurtleff SA. Genetic analysis of patients with defects in early B-cell development. *Immunol Rev* 2005;**203**:216–34.

9 van Zelm MC, Geertsema C, Nieuwenhuis N, de Ridder D, Conley ME, Schiff C, Tezcan I, Bernatowska E, Hartwig NG, Sanders EAM, Litzman J, Kondratenko I, van Dongen JJM, van der Burg M. Gross deletions involving IGHM, BTK, or Artemis: a model for genomic lesions mediated by transposable elements. *Am J Hum Genet* 2008;**82**:320–32.

10 Jo E-K, Wang Y, Kanegane H, Futatani T, Song C-H, Park J-K, Kim JS, Kim DS, Ahn K-M, Lee S-I, Park HJ, Hahn YS, Lee J-H, Miyawaki T. Identification of mutations in the Bruton's tyrosine kinase gene, including a novel genomic rearrangements resulting in large deletion, in Korean X-linked agammaglobulinemia patients. *J Hum Genet* 2003;**48**:322–6.

11 Richter D, Conley ME, Rohrer J, Myers LA, Zahradka K, Kelecić J, Sertić J, Stavljenić-Rukavina A. A contiguous deletion syndrome of X-linked agammaglobulinemia and sensorineural deafness. *Pediatr Allergy Immunol* 2001;**12**:107–11.

12 Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;**16**:1215.

13 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

14 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.

15 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

16 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8.

17 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

18 Bousfiha A, Jeddane L, Picard C, Al-Herz W, Ailal F, Chatila T, Cunningham-Rundles C, Etzioni A, Franco JL, Holland SM, Klein C, Morio T, Ochs HD, Oksenhendler E, Puck J, Torgerson TR, Casanova J-L, Sullivan KE, Tangye SG. Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification. *J Clin Immunol* 2020;**40**:66–81.

19 Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012;**20**:490–7.

20 Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–94.

21 Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**:R84.

22 Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* 2015;**12**:966–8.

23 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6.

24 Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet* 2012;**8**:e1002453.

25 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.

26 Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* 2008;**9**:540.

27 Bruton OC. Agammaglobulinemia. *Pediatrics* 1952;**9**:722–8.

28 Soresina A, Nacinovich R, Bomba M, Cassani M, Molinaro A, Sciotto A, Martino S, Cardinale F, De Mattia D, Putti C, Dellepiane RM, Felici L, Parrinello G, Neri F, Plebani A, Italian Network for Primary Immunodeficiencies. The quality of life of children and adolescents with X-linked agammaglobulinemia. *J Clin Immunol* 2009;**29**:501–7.

29 Zhang Y-N, Gao Y-Y, Yang S-D, Cao B-B, Zheng K-L, Wei P, Chen L-F, Chen W-X. Delayed diagnosis of X-linked agammaglobulinaemia in a boy with recurrent meningitis. *BMC Neurol* 2019;**19**:320.

30 Preece K, Lear G. X-linked Agammaglobulinemia With Normal Immunoglobulin and Near-Normal Vaccine Seroconversion. *Pediatrics* 2015;**136**:e1621–4.

31 Lim L-M, Chang J-M, Wang I-F, Chang W-C, Hwang D-Y, Chen H-C. Atypical X-linked agammaglobulinaemia caused by a novel BTK mutation in a selective immunoglobulin M deficiency patient. *BMC Pediatr* 2013;**13**:150.

32 Sheppard S, Biswas S, Li MH, Jayaraman V, Slack I, Romasko EJ, Sasson A, Brunton J, Rajagopalan R, Sarmady M, Abrudan JL, Jairam S, DeChene ET, Ying X, Choi J, Wilkens A, Raible SE, Scarano MI, Santani A, Pennington JW, Luo M, Conlin LK, Devkota B, Dulik MC, Spinner NB, Krantz ID. Utility and limitations of exome sequencing as a genetic diagnostic tool for children with hearing loss. *Genet Med* 2018;**20**:1663–76.

33 Biesecker LG, Shianna KV, Mullikin JC. Exome sequencing: the expert view. *Genome Biol* 2011;**12**:128.

34 Chou J, Ohsumi TK, Geha RS. Use of whole exome and genome sequencing in the identification of genetic causes of primary immunodeficiencies. *Curr Opin Allergy Clin Immunol* 2012;**12**:623–8.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

35 Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova J-L, Abel L. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* 2015;**112**:5473–8.

36 Thaventhiran JED, Allen HL, Burren OS, Rae W, Greene D, Staples E, Zhang Z, Farmery JHR, Simeoni I, Rivers E, Maimaris J, Penkett CJ, Stephens J, Deevi SVV, Sanchis-Juan A, Gleadall NS, Thomas MJ, Sargur RB, Gordins P, Baxendale HE, Brown M, Tuijnenburg P, Worth A, Hanson S, Linger RJ, Buckland MS, Rayner-Matthews PJ, Gilmour KC, Samarghitean C, Seneviratne SL, Sansom DM, Lynch AG, Megy K, Ellinghaus E, Ellinghaus D, Jorgensen SF, Karlsen TH, Stirrups KE, Cutler AJ, Kumararatne DS, Chandra A, Edgar JDM, Herwadkar A, Cooper N, Grigoriadou S, Huissoon AP, Goddard S, Jolles S, Schuetz C, Boschann F, Lyons PA, Hurles ME, Savic S, Burns SO, Kuijpers TW, Turro E, Ouwehand WH, Thrasher AJ, Smith KGC. Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature* 2020;**583**:90–5.

37 Guo X, Delio M, Haque N, Castellanos R, Hestand MS, Vermeesch JR, Morrow BE, Zheng D. Variant discovery and breakpoint region prediction for studying the human 22q11.2 deletion using BAC clone and whole genome sequencing analysis. *Hum Mol Genet* 2016;**25**:3754–67.

38 Mohamed AJ, Yu L, Bäckesjö C-M, Vargas L, Faryal R, Aints A, Christensson B, Berglöf A, Vihinen M, Nore BF, Smith CIE. Bruton's tyrosine kinase (Btk): function, regulation, and transformation with special emphasis on the PH domain. *Immunol Rev* 2009;**228**:58–73.

39 Fukuda M, Kojima T, Kabayama H, Mikoshiba K. Mutation of the pleckstrin homology domain of Bruton's tyrosine kinase in immunodeficiency impaired inositol 1,3,4,5-tetrakisphosphate binding capacity. *J Biol Chem* 1996;**271**:30303–6.

40 Yip KL, Chan SY, Ip WK, Lau YL. Bruton's tyrosine kinase mutations in 8 Chinese families with X-linked agammaglobulinemia. *Hum Mutat* 2000;**15**:385.

41 Ochs MD Dr.med HD, Smith PhD CIE, Puck MD JM, editors. *Primary Immunodeficiency Diseases*. Oxford University Press 2013.

42 Teocchi MA, Domingues Ramalho V, Abramczuk BM, D'Souza-Li L, Santos Vilela MM.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

BTK mutations selectively regulate BTK expression and upregulate monocyte XBP1 mRNA in XLA patients. *Immun Inflamm Dis* 2015;**3**:171–81.

43 Wright PF, Hatch MH, Kasselberg AG, Lowry SP, Wadlington WB, Karzon DT. Vaccine-associated poliomyelitis in a child with sex-linked agammaglobulinemia. *J Pediatr* 1977;**91**:408–12.

44 van Maarle MC, Stouthard MEA, Bonsel GJ. Quality of life in a family based genetic cascade screening programme for familial hypercholesterolaemia: a longitudinal study among participants. *J Med Genet* 2003;**40**:e3.

45 Ahmed S, Saleem M, Modell B, Petrou M. Screening extended families for genetic hemoglobin disorders in Pakistan. *N Engl J Med* 2002;**347**:1162–8.

46 Madrigal I, Rodríguez-Revenga L, Badenas C, Sánchez A, Martinez F, Fernandez I, Fernández-Burriel M, Milà M. MLPA as first screening method for the detection of microduplications and microdeletions in patients with X-linked mental retardation. *Genet Med* 2007;**9**:117–22.

47 Lalic T, Vossen RHAM, Coffa J, Schouten JP, Guc-Scekic M, Radivojevic D, Djurisic M, Breuning MH, White SJ, den Dunnen JT. Deletion and duplication screening in the DMD gene using MLPA. *Eur J Hum Genet* 2005;**13**:1231–4.

48 Vellaichamy Swaminathan V, Uppuluri R, Patel S, Melarcode Ramanan K, Ravichandran N, Jayakumar I, Vaidhyanathan L, Raj R. Treosulfan-based reduced toxicity hematopoietic stem cell transplantation in X-linked agammaglobulinemia: A cost-effective alternative to long-term immunoglobulin replacement in developing countries. *Pediatr Transplant* 2020;**24**:e13625.

49 Ikegame K, Imai K, Yamashita M, Hoshino A, Kanegane H, Morio T, Kaida K, Inoue T, Soma T, Tamaki H, Okada M, Ogawa H. Allogeneic stem cell transplantation for X-linked agammaglobulinemia using reduced intensity conditioning as a model of the reconstitution of humoral immunity. *J Hematol Oncol* 2016;**9**:9.

50 El-Sayed ZA, Abramova I, Aldave JC, Al-Herz W, Bezrodnik L, Boukari R, Bousfiha AA, Cancrini C, Condino-Neto A, Dbaibo G, Derfalvi B, Dogu F, Edgar JDM, Eley B, El-Owaidy

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

RH, Espinosa-Padilla SE, Galal N, Haerynck F, Hanna-Wakim R, Hossny E, Ikinciogullari A, Kamal E, Kanegane H, Kechout N, Lau YL, Morio T, Moschese V, Neves JF, Ouederni M, Paganelli R, Paris K, Pignata C, Plebani A, Qamar FN, Qureshi S, Radhakrishnan N, Rezaei N, Rosario N, Routes J, Sanchez B, Sediva A, Seppanen MR, Serrano EG, Shcherbina A, Singh S, Siniah S, Spadaro G, Tang M, Vinet AM, Volokha A, Sullivan KE. X-linked agammaglobulinemia (XLA):Phenotype, diagnosis, and therapeutic challenges around the world. *World Allergy Organ J* 2019;**12**:100018.

51 Rudilla F, Franco-Jarava C, Martínez-Gallo M, Garcia-Prat M, Martín-Nalda A, Rivière J, Aguiló-Cucurull A, Mongay L, Vidal F, Solanich X, Irastorza I, Santos-Pérez JL, Tercedor Sánchez J, Cuscó I, Serra C, Baz-Redón N, Fernández-Cancio M, Carreras C, Vagace JM, Garcia-Patos V, Pujol-Borrell R, Soler-Palacín P, Colobran R. Expanding the Clinical and Genetic Spectra of Primary Immunodeficiency-Related Disorders With Clinical Exome Sequencing: Expected and Unexpected Findings. *Front Immunol* 2019;**10**:2325.

Acknowledgements

The work was supported by Council of Scientific and Industrial Research (CSIR) India through grant number MLP1801 (RareGen-CSIR India), Science & Engineering Research Board (SERB) through grant number EMR/2016/006828/HS (SERB, DST) and Foundation for Primary Immunodeficiency Diseases (FPID, USA).

We also thank Pulukool Sandhya, Disha Sharma, Bani Jolly, Mukta Poojary, and Afra Shamnath for suggestions which enriched the manuscript.

Conflict of Interest

Authors declare no conflict of interest.

Data Availability:

The sequencing data in the study will be provided upon request, as it could be potentially identifiable and contains patient sensitive information. Access to the sequencing data could be requested by mailing to Dr. Jyoti Yadav (j.yadav@igib.res.in) who is the convener of the Institutional Human Ethics Committee of CSIR-Institute of Genomics and Integrative Biology. Similarly, the pedigree of the family and descriptive patient clinical information will

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

be provided upon request to the corresponding author Dr. Vinod Scaria (vinods@igib.in) Principal Scientist of CSIR-Institute of Genomics and Integrative Biology.

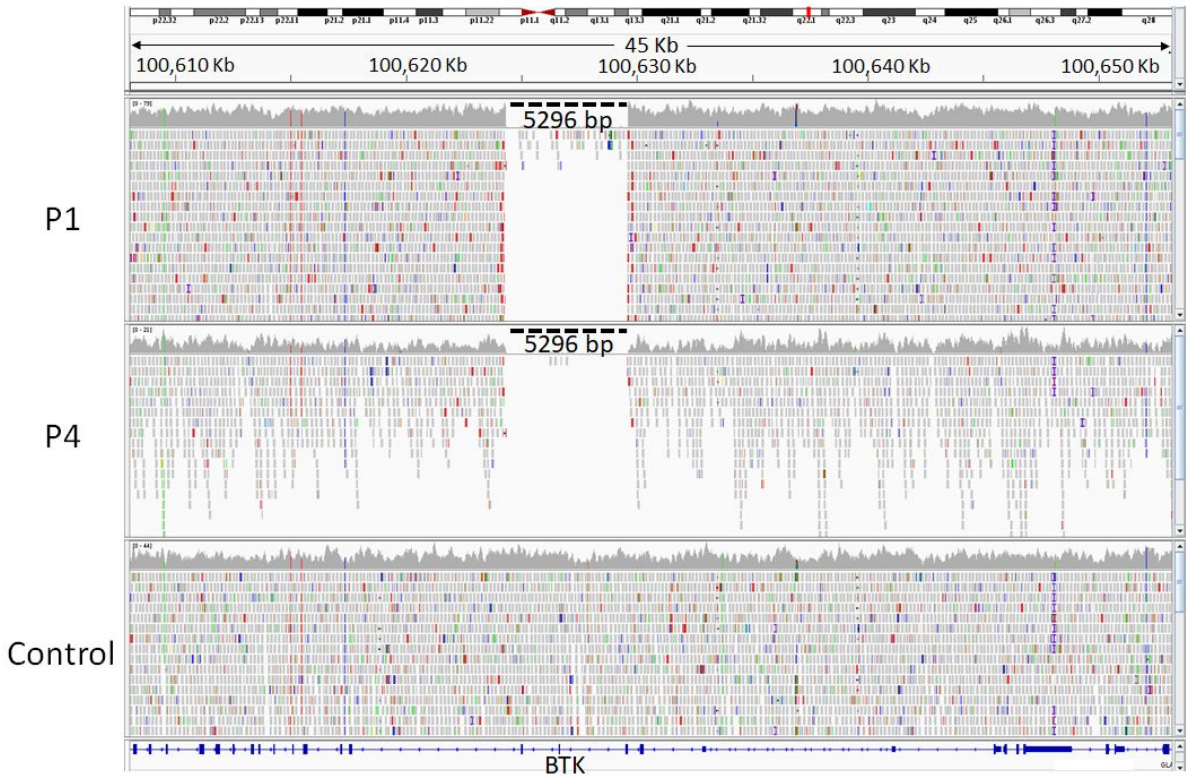

Figure 1: Identification of putative large deletion in whole genome data in P1 and P4: The figure represents the visualization of sequencing reads aligned on the coding region of the *BTK* gene.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

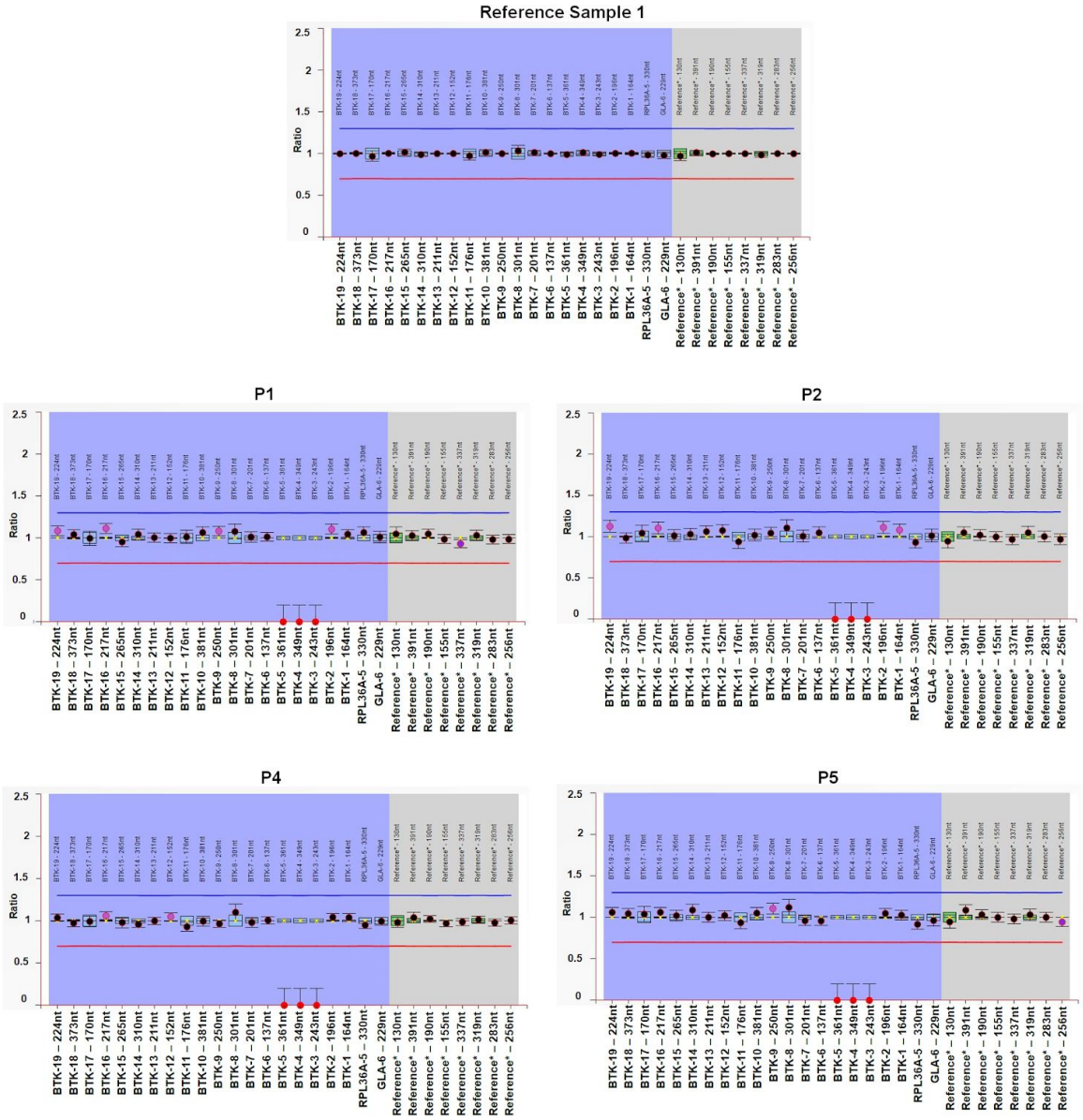

Figure 2: Multiplex ligation dependent amplification (MLPA) assay for detection of deletion encompassing 3-5 exons of the *BTK* gene: Representation of ratio charts generated using the Coffalyser.net software. (A) Reference sample depicts the normal copies of the BTK gene and other samples (B) Sample P1 (C) Sample P2 (D) Sample P4 (E) Sample P5 are all males and depict probe ratio of zero for 3-5 exons of the BTK gene which lies on X-chromosome, i.e. they have hemizygous deletion. For all the ratio charts, the

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

longitudinal axis (X-axis) represents the final ratios after inter and intra normalisations of the probe ratios. Horizontal axis (Y-axis) represents the probe names alongwith the length in the order (Both axis title have been manually enlarged for clarity). The blue and red horizontal lines depict the arbitrary borders of ratio 1.3 and 0.7 respectively. The black and red dots denote the final ratio obtained for each of the probes and the vertical bars represent the 95% confidence range for each probe.

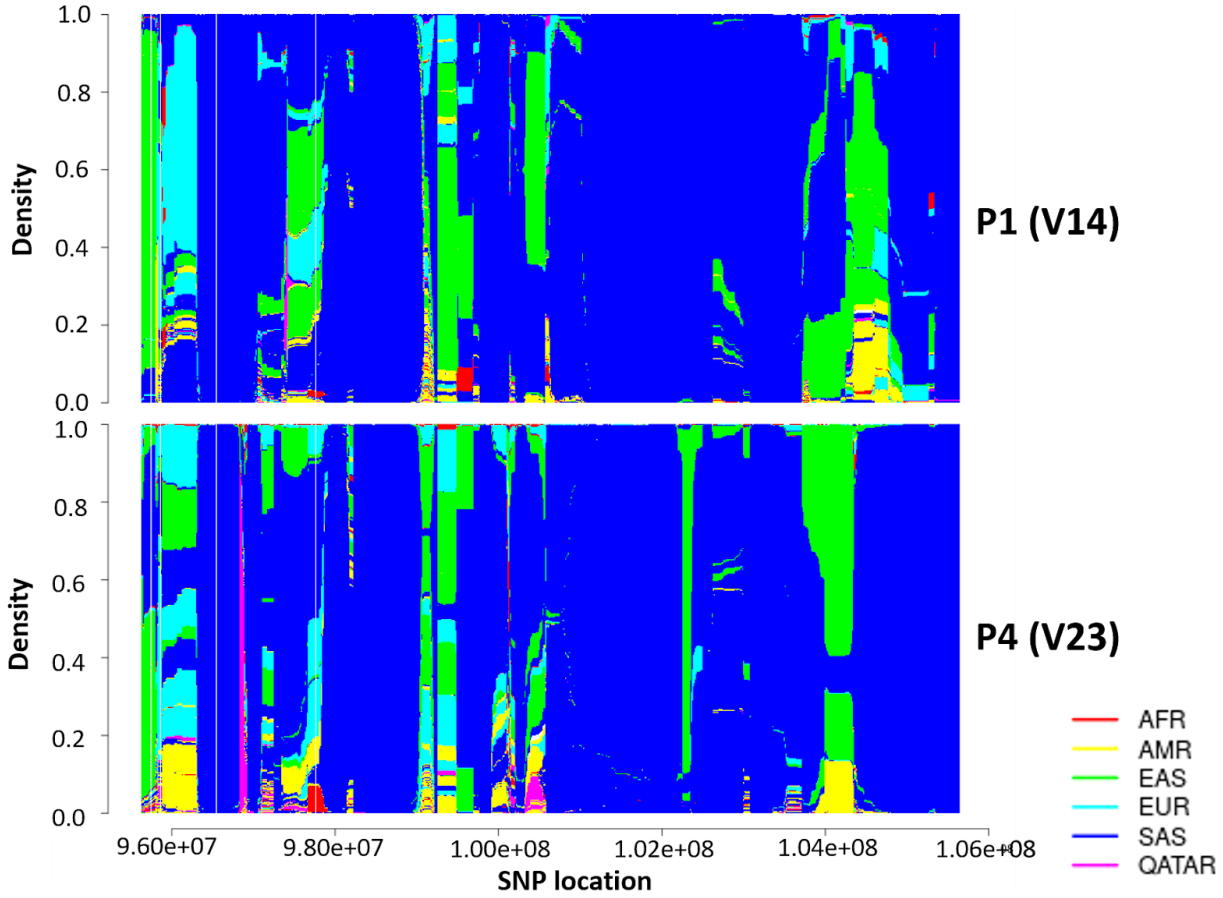

Figure 3: Haplotype ancestry prediction with the global population using fineSTRUCTURE. South Asian ancestry predicted the haplotype for locus 5MB upstream and 5MB downstream to loci chrX:100,624,323-100,629,619 (hg19/GRCh37) of two affected patients (V14 and V23) with 2504 individuals of five major populations (AFR-African, AMR- American, EAS- East Asian, EUR-European, and SAS-South Asian) of 1000 Genome Project and 44 individuals of Qatari ancestry.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Tables

Clinical and immunological characteristics	**P1**	**P2**	**P3**	**P4**	**P5**
Age at onset	0-5 years	0-5 years	0-5 years	0-5 years	0-5 years
Age at diagnosis	6-10 years	0-5 years	0-5 years	0-5 years	6-10 years
Type of infections	Pneumonia, Pyoderma, Pyogenic meningitis	Pneumonia, Acute suppurative otitis media, Diarrhea, Oral thrush	Pneumonia Diarrhea Pyoderma	Diarrhea	Pneumonia, Pyogenic meningitis Diarrhea, pyoderma
Absent / atrophic tonsils	yes	yes	yes	yes	yes
Sibling death	yes	no	yes	yes	no
No. of hospitalizations before diagnosis	2	>15	1	2	>10
No. of PICU admissions	none	2	1	none	1
Immunoglobulin assay	IgG low, IgA low, IgM low	IgG low, IgA low, IgM normal	IgG low, IgA normal, IgM normal	IgG low, Ig A low, IgM normal	IgG low, IgA low, IgM normal
CD 19 count at diagnosis (N) In cells/mm^3	0 (91-610)	0 (430-3300)	8 (430-3300)	10 (430-3300)	10 (91-610)
CD 3 count at diagnosis (N) in cells/mm^3	6458 (570-2400)	5475 (1460-5440)	6400 (1460-5440)	4169 (1900-5900)	9553 (570-2400)
CD 56 count at diagnosis (N) in	355 (78-470)	7.0175 (80-340)	160 (80-340)	661 (160-950)	528 (78-470)

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

cells/mm^3					
Onset to IVIG start	Not on regular IVIG	1 year, 2 months	11 months	1 month	Not on regular IVIG
Complications	Bronchiectasis	None	None	None	Bronchiectasis, Arthritis
Outcome	Stunted growth,	Stunted growth, Cured – HSCT from matched sibling donor (twin) at 6.5 years	Stunted growth Died	Cured- HSCT from matched sibling donor at 2 years	Stunted growth and delayed puberty

Table 1: Clinical and Immunological features of patients with XLA in the present cohort

Data	**P1**	**P4**
Total Structural Variants	8,925	29,638
Common homozygous Variants	768	
PID genetic variants (454 gene)	8	
Phenotype associated variant	1	
Structural Variant	chrX:100,624,323-100,629,619 (5296 bp deletion)	

Table 2: Whole genome sequencing structural variant analysis for P1 and P4.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.05.20200949; this version posted October 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

Supplementary Material

Supplementary Tables

Supplementary Table 1: Whole exome sequencing data summary for each patient.

Supplementary Table 2: Whole genome sequencing data summary for P1 and P4.

Supplementary Table 3: MLPA copy number ratios for each of the reference and test probes of the proband and all the extended family members tested.

Supplementary Figures:

Supplementary Figure 1: Visualization of putative large deletion encompassing *BTK* gene exon 3-5 of whole exome data of patients P1, P2, P4, P5, and control sample.

Supplementary Figure 2: MLPA for detection of deletion encompassing 3-5 exons of the *BTK* gene in the additional family members of the proband: Representation of ratio charts generated using the Coffalyser.net software. Longitudinal axis represents the final ratios after inter and intra normalisations of the probe ratios. Horizontal axis represents the probe names alongwith the length (The axis titles have been manually enlarged for clarity). The blue and red horizontal lines depict the arbitrary borders of ratio 1.3 and 0.7 respectively. The black and red dots denote the final ratio obtained for each of the probes and the vertical bars represent the 95% confidence range for each probe. The test probes are of *BTK* gene and the rest are the reference probes.

Supplementary Figure 3: South Asian ancestry predicted for locus A) 50MB upstream and 5KB downstream, B) 500KB upstream and 500KB downstream, and C) 50KB upstream and 50KB downstream to loci chrX:100,624,323-100,629,619 (hg19/GRCh37) of two affected first cousins (V14 and V23) with 2504 individuals of five major populations (AFR-African, AMR- American, EAS- East Asian, EUR-European, and SAS-South Asian) of 1000 Genome Project and 44 individuals of Qatar ancestry.