Rare genetic diseases variant calling prediction

Joel Frayle Moreno. *, Francesco Mazza. †
Scuola di Ingegneria Industriale e dell'Informazione, Politecnico di Milano. Milan, Italy

***Abstract*—In this study, we performed a comprehensive analysis of whole-exome sequencing data from multiple familial trios to identify potential causal variants underlying rare genetic disorders. High-quality read alignment was achieved using Bowtie2, followed by variant calling with FreeBayes, employing stringent filters to ensure robust and reliable detection. Variant prioritization focused on exonic regions, guided by a BED file including padded exon coordinates to capture splice-affecting events. Inheritance models—autosomal dominant and autosomal recessive—were applied to filter variants based on the genotypes of parents and offspring. Resulting variant call files (VCFs) were annotated using Ensembl's Variant Effect Predictor (VEP), incorporating population allele frequencies, predicted functional impact, and clinical significance. In nine out of ten cases, we identified high-confidence variants consistent with known disease phenotypes such as Fanconi anemia, autosomal dominant polycystic kidney disease, and Rubinstein-Taybi syndrome. One case remained inconclusive due to the detection of a common variant with low predicted impact, highlighting the need for further investigation. This pipeline demonstrates the power of integrated computational approaches in rare disease diagnosis through trio-based exome analysis.**

Keywords: whole exome sequencing, variant calling, FreeBayes, VEP, trio analysis, rare diseases, inheritance models, autosomal dominant, autosomal recessive.

I. Introduction

Rare genetic diseases are a heterogeneous group of disorders that collectively affect hundreds of millions of people worldwide. Although each individual condition may impact only a small number of individuals—often fewer than 1 in 2,000 people—they represent a significant global health burden due to their sheer number, complexity, and the challenges associated with accurate diagnosis and management. To date, nearly 10,000 rare diseases have been identified, the majority of which have a genetic origin and manifest early in life, particularly in pediatric populations [1], [2].

The World Health Organization (WHO) estimates that over 6,000 rare diseases collectively affect approximately 1 in 2,000 individuals. Despite advances in genomic medicine, diagnosing rare diseases remains a formidable task for healthcare systems. One of the major challenges lies in the low prevalence of each disease, which reduces clinical awareness and hinders the establishment of well-defined diagnostic criteria. Furthermore, the wide variability in clinical presentation, even among patients with the same disorder, complicates efforts to reach a precise genetic diagnosis [3].

Among the current strategies employed for diagnosing rare genetic conditions, whole genome sequencing (WGS) and whole exome sequencing (WES) have emerged as powerful tools. In particular, WES focuses on the protein-coding regions of the genome—known as exons—which harbor the majority of pathogenic mutations associated with Mendelian disorders. The identification of single nucleotide variants (SNVs) within these regions is a critical step in the diagnostic pipeline, as these point mutations can disrupt gene function and lead to disease phenotypes [4].

Rare genetic diseases following autosomal inheritance patterns can generally be classified into two major categories: autosomal recessive (AR) and autosomal dominant (AD). In AR conditions, a patient must inherit two pathogenic alleles—one from each parent—in order to express the disease. In contrast, AD disorders require only a single pathogenic allele to manifest, meaning that a patient can develop the disease even if both parents are asymptomatic carriers of the non-mutated allele. Furthermore, in AD diseases, it is possible for the causative mutation to arise de novo, as a result of spontaneous errors during DNA replication or early embryonic development [5].

In this study, we describe a bioinformatics pipeline for the detection of rare genetic diseases through variant calling in whole genome data. Leveraging a Unix-based environment and open-source tools such as Bowtie2 for alignment, SAMtools for sequence manipulation, and FreeBayes for variant calling, we demonstrate an efficient approach for the analysis and identification of causative mutations. This pipeline is designed to be both scalable and reproducible, facilitating its adoption in research and clinical settings aimed at uncovering the genetic basis of rare diseases [6].

II. Methodology

First, for each case, we focused on analyzing the sequence quality to ensure that the genomic data was sufficiently reliable for subsequent analyses. This was carried out using the FastQC tool for each individual file, as shown in lines 50 to 63 (Available on the Github repository). Once all FastQC analyses were completed, we utilized the MultiQC tool [7] to compile the individual FastQC reports into a single comprehensive report. This allowed us to assess the quality of all datasets within each case collectively yet individually. This step was performed in line 67.

Subsequently, between lines 87 and 98, alignment was performed using the Bowtie2 tool, which enables alignment against a reference model. Each individual's data (mother, father, and child) was aligned to the "uni" index of Chromosome 16 (chr16). At this step, reference identifiers were also adjusted to clearly distinguish between the child, mother, and father. Within the same line of code, to optimize storage space and streamline subsequent processes, the resulting SAM files were converted to binary BAM format using the 'samtools view -Sb' command. The BAM files were then sorted by genomic position using 'samtools sort', facilitating efficient access for subsequent analyses.

After obtaining all alignments, indices were generated using ‘samtools index‘, as illustrated in lines 104 to 107. This indexing allowed optimized access to specific genomic regions without needing to load the entire file.

Finally, variant comparison analyses were conducted using the FreeBayes tool, comparing chromosome 16 sequences between the mother, father, and child for each case. Thus, each individual’s genome was compared against chr16 and, consequently, among each other. As indicated in line 113, several filters were implemented to guarantee result quality and robustness, including a minimum mapping quality of 20, a variant support threshold of 5, a minimum base quality of 10, and a minimum coverage of 10 reads. Additionally, the analysis was restricted to exon regions using the BED file exons16Padded_sorted.bed (option -t). This filter focuses variant calling specifically on exonic sequences, biologically justified by the higher likelihood of exonic variants to directly affect protein structure and function, thereby increasing their potential clinical relevance and interpretability.

Additional filtering was then applied based on whether the condition was autosomal recessive or dominant. For recessive diseases transmitted from heterozygous parents to a homozygous child, and considering the VCF file order provided by FreeBayes (mother, father, child), the genotype pattern must follow ’GT[2]=”1/1” && GT[0]=”0/1” && GT[1]=”0/1”’, signifying heterozygous parents and a homozygous affected child. Conversely, for autosomal dominant conditions—given that neither parent has the variant—the genotype pattern should be ’GT[0]=”0/0” && GT[1]=”0/0” && GT[2]=”0/1”’, indicating a novel mutation in the child. Furthermore, a quality filter greater than 20 was applied to further reduce candidate variants. As shown in lines 117 to 125, this filtering process produced a VCF file containing chromosome 16 positions, allele distributions, and additional relevant information for each candidate recessive disorder.

Lastly, we used the Variant Effect Predictor (VEP) tool to cross-reference these candidate variants with literature and existing databases. VEP analysis provided detailed annotations, including predicted impacts, clinical classifications (benign, pathogenic, etc.), and phenotype-related literature. Consequently, we could identify rare genetic diseases among our candidates, confirming their presence on the child’s chromosome.

III. Results

Plot 3 shows the general quality of all our case files, highlighting the mean quality value for each base position in the read.

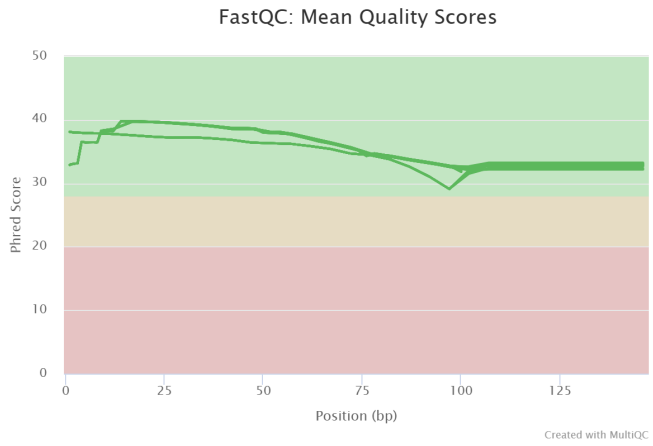

Figure 1: Per-base sequence quality plot showing the mean quality value across each base position in the read. The values consistently range between 30 and 40.

the fact that the quality is in the green part, between 30 and 40, indicates high sequencing quality. These results support the reliability of downstream analyses from a quality standpoint.

Exome sequencing targets only protein-coding regions, which are about 1–2% of the genome.

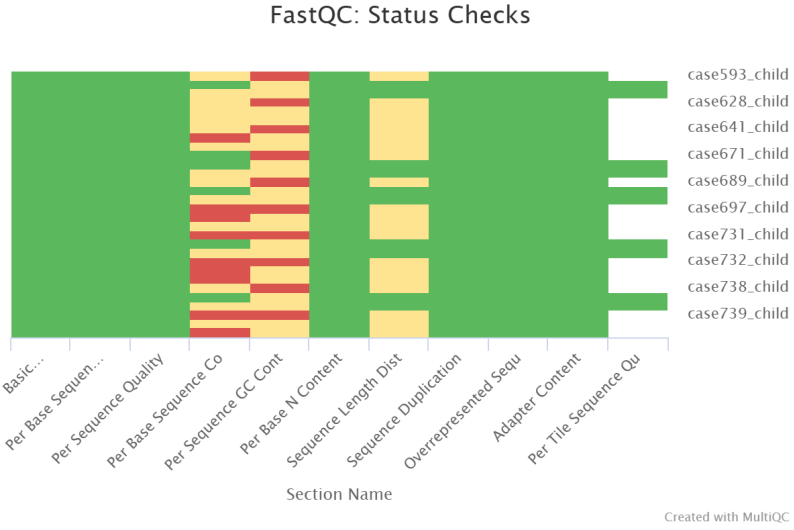

Figure 2: Status Check: Summary of FastQC quality checks. Green indicates good quality, yellow flags potential issues, and red marks poor-quality.

The GC percentage is not high; the protein-coding regions don’t evenly represent the genome’s GC distribution. Exome capture kits can have biases: they may under-represent GC-rich or highly repetitive regions, depending on probe design [8]. As a result, exome sequencing data often show: Lower GC content than whole genome data. Skewed or non-uniform GC distribution, depending on capture efficiency.

The Bowtie2 tool generates binary alignment files (.bam), which are not intended for direct visual inspection due to their binary format. These .bam files represent an intermediate step that facilitates subsequent analytical processes, rather than a final interpretable output. Similarly, indexing files generated by

samtools (.bam.bai extension) optimize processing speed for subsequent analyses, as previously described in the methodology.

Subsequently, variant analyses were conducted using FreeBayes, producing variant call format (VCF) files named following the structure Triocase number.vcf. As illustrated in Figure 3 using case 593 as an example, each VCF file begins with a header section containing metadata that specifies the analysis parameters and tools used, followed by a matrix where each row represents an identified variant. The columns of this matrix include information such as chromosome, exact genomic position, reference and alternate alleles, variant calling quality (QUAL), and various supporting technical statistics. Additionally, individual genotypes for each trio member (father, mother, child) are specified, alongside key metrics like read depth (DP), allelic depth (AD), and genotype likelihoods (GL). This structure facilitates precise variant characterization and enables filtering based on criteria consistent with distinct inheritance patterns. General statistics, such as total variant count, proportions of heterozygous and homozygous variants, and overall distribution of variant call qualities, were extracted for each case to facilitate comparative analyses.

Figure 3 demonstrates the typical structure of a VCF file, highlighting the header and variant matrix detailing chromosome, position, reference and alternate alleles, QUAL scores, and specific parameters for each sample.

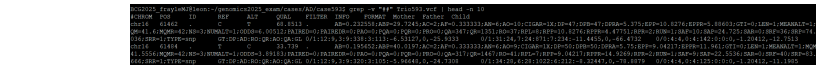

Figure 3: General structure of a VCF file. The header section is followed by a matrix detailing variants identified, with columns specifying chromosome, genomic position, reference and alternate alleles, variant calling quality (QUAL), and specific parameters for each analyzed sample.

The total number of variants identified per trio was quantified: cases 593, 628, 641, 689, and 738 yielded 7,657 variants each; cases 671, 697, 731, and 739 presented 7,656 variants each; and case 732 yielded 7,655 variants. This quantification was conducted using the procedure described in line 134 of the analysis script. After applying the filtering criteria described in the methodology, the remaining candidate variants potentially associated with dominant or recessive genetic disorders are summarized in Table I. These filtered variants were subsequently analyzed with the Variant Effect Predictor (VEP).

Table I summarizes the final number of candidate variants identified per case after filtering:

Table I: Total number of candidate variants identified per case.

Case	Inheritance (AD/AR)	Number of Variants
case593	AD	16
case628	AR	183
case641	AR	183
case671	AD	15
case689	AD	16
case697	AR	183
case731	AD	16
case732	AR	183
case738	AD	16
case739	AR	183

The candidate variants analyzed through VEP allowed the identification of potential effects of these variations on the child. After applying further filtering criteria, focusing on variants of high impact, low population frequency, and pathogenic clinical significance, final diagnoses and associated variant details are presented in Table II.

IV. Results analysis

The final results following the analysis performed with the Variant Effect Predictor (VEP) are presented in Table II. For each trio, we report the candidate variant, including its genomic location, the altered allele, the predicted consequence and functional impact, the affected gene, the allele frequency in the general population, and the associated phenotype.

Importantly, all cases—except one (case 671)—present variants with a predicted HIGH impact according to VEP, suggesting a severe effect on protein structure or function, such as premature stop codons (stop gained) or frameshift events. These mutation types are widely recognized as deleterious, which aligns with the hypothesis that these variants may underlie monogenic disease phenotypes in the affected children.

In the autosomal dominant (AD) group, cases 593 and 738 both presented stop gained mutations in the PKD1 gene, which is consistent with Autosomal Dominant Polycystic Kidney Disease (ADPKD). ADPKD is a frequent monogenic disorder characterized by the progressive development of renal cysts leading to chronic kidney failure, and truncating mutations in PKD1 are well-known causal variants [9], [10]. Similarly, case 689 showed a stop gained mutation in CREBBP, a gene linked to Rubinstein-Taybi syndrome, a condition marked by intellectual disability, broad thumbs, and characteristic facial features [11], [12].

Case 731 presented a high-impact stop gained mutation in the SALL1 gene, which is strongly associated with Townes-Brocks syndrome, a rare disorder involving renal malformations, imperforate anus, and limb anomalies [13].

Case 671, however, presented a distinct scenario. A splice polypyrimidine variant of LOW impact was detected in the TSC2 gene. While TSC2 is involved in tuberous sclerosis complex, the variant found has an allele frequency of 3.43%, which is inconsistent with the expected frequency of rare genetic disorders, typically below 0.1% in the general population. Moreover, the predicted functional effect is minimal. Therefore, this variant does not provide statistically or biologically sufficient evidence to support a definitive diagnosis, and this case remains inconclusive pending further investigation.

In the autosomal recessive (AR) group, there is a striking convergence on the FANCA gene in four cases (628, 641, 732, 739), all of which presented stop gained or frameshift variants. This gene is associated with Fanconi anemia (FA), a genetic disorder characterized by bone marrow failure, congenital malformations, and cancer predisposition [14], [15]. The identified mutations align with known pathogenic mechanisms in FA.

Case 697 revealed a frameshift mutation in the CIITA gene, which is the master regulator of MHC class II gene expression. Mutations in this gene cause Bare Lymphocyte Syndrome,

Table II: Final VEP results and genetic diagnoses per case.

Case	Location	Allele	Consequence	Impact	Gene (Symbol)	Allelic Frequency	Associated Phenotype
case593	chr16:2140777-2140777	T	stop gained	HIGH	PKD1	-	AD polycystic kidney disease
case628	chr16:89815164-89815174	-	frameshift variant	HIGH	FANCA	-	Fanconi anemia
case641	chr16:89858892-89858892	C	stop gained	HIGH	FANCA	-	Fanconi anemia
case671	chr16:2115506-2115506	T	splice polypyrimidine	LOW	TSC2	0.0343	Adult hepatocellular carcinoma
case689	chr16:3789608-3789608	C	stop gained	HIGH	CREBBP	-	Rubinstein-Taybi syndrome 1
case697	chr16:10996514-10996519	-	frameshift variant	HIGH	CIITA	-	MHC class II deficiency
case731	chr16:51174877-51174877	T	stop gained	HIGH	SALL1	-	Townes-Brocks syndrome
case732	chr16:89858416-89858416	A	stop gained	HIGH	FANCA	-	Fanconi anemia
case738	chr16:2142182-2142182	C	stop gained	HIGH	PKD1	-	AD polycystic kidney disease
case739	chr16:89858357-89858361	-	frameshift variant	HIGH	FANCA	-	Fanconi anemia

Type II (MHC Class II Deficiency), a primary immunodeficiency with early-onset infections and poor immune responses [16]. The result is entirely consistent with the clinical picture expected for this condition.

V. Coverage

In order to understand how it's possible to visualzie the data on the genome browser, here in Figure 4 there is an example with case628 Trio coverage. Across the 131 bp window on chr16, read-depth in the proband plunges to zero over a roughly 40 bp central interval while both parents retain only a faint residual signal (5–10 × versus 30–40 × in the flanking regions), a pattern diagnostic of a homozygous deletion in the child with each parent as a heterozygous carrier; this deletion precisely co-localizes with the "GGAGGAG" indel breakpoints reported in Trio628.vcf and overlays multiple GENCODE V45lift37 transcripts of the FANCA gene, implicating a loss of FANCA coding sequence consistent with a recessive Fanconi anemia–causing lesion.

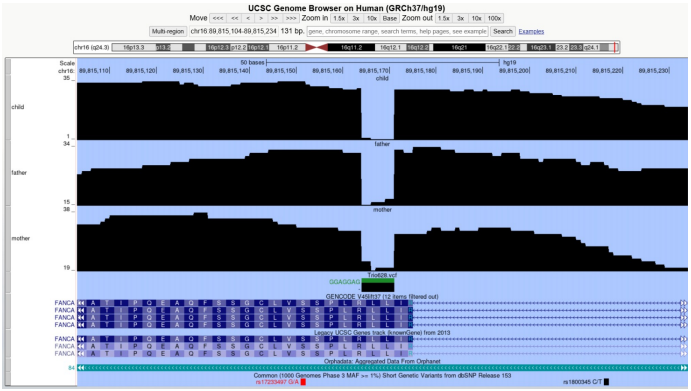

Figure 4: General UCSC Genome Browser view of the FANCA gene region (chr16:89,815,104–89,815,234, GRCh37/hg19) in a trio (child, father, mother).

VI. Conclutions

In conclusion, the variant filtering strategy using VEP allowed us to narrow down thousands of initial candidates to a small number of high-confidence variants per trio. The integration of functional predictions, population frequency data, and genotype-phenotype consistency enabled us to define a likely genetic diagnosis in nine out of ten cases. The one remaining case (case 671) requires extended genetic or functional exploration to clarify its etiology.

References

[1] K. M. Boycott, T. Hartley, L. G. Biesecker, *et al.*, *A Diagnosis for All Rare Genetic Diseases: The Horizon and the Next Frontiers*, Mar. 2019. DOI: 10.1016/j.cell.2019.02.040.

[2] I. Rashid, P. S, S. Moharir, and R. Mishra, *GenTIGS: A database empowering research and clinical insights on rare genetic disorders with an Indian perspective*, Apr. 2025. DOI: 10.1101/2025.04.01.25325014. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2025.04.01.25325014.

[3] H. Han, G. H. Seo, S.-I. Hyun, *et al.*, "Exome sequencing of 18,994 ethnically diverse patients with suspected rare Mendelian disorders," *npj Genomic Medicine 2025 10:1*, vol. 10, no. 1, pp. 1–9, Jan. 2025, ISSN: 2056-7944. DOI: 10.1038/s41525-024-00455-3. [Online]. Available: https://www.nature.com/articles/s41525-024-00455-3.

[4] G. Lai, Q. Gu, Z. Lai, H. Chen, J. Chen, and J. Huang, "The application of whole-exome sequencing in the early diagnosis of rare genetic diseases in children: a study from Southeastern China," *Frontiers in Pediatrics*, vol. 12, p. 1 448 895, 2024, ISSN: 22962360. DOI: 10.3389/FPED.2024.1448895/FULL. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11493614/.

[5] K. Slaba, P. Pokorna, R. Jugas, *et al.*, "Diagnostic efficacy and clinical utility of whole-exome sequencing in Czech pediatric patients with rare and undiagnosed diseases," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–11, Nov. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-79872-4. [Online]. Available: https://www.nature.com/articles/s41598-024-79872-4.

[6] R. Pereira, J. Oliveira, and M. Sousa, "Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics," *Journal of Clinical Medicine*, vol. 9, no. 1, p. 132, Jan. 2020, ISSN: 20770383. DOI: 10.3390/JCM9010132. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7019349/.

[7] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, "Multiqc: Summarize analysis results for multiple tools and samples in a single report," *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Jun. 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw354. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btw354.

[8] M. Dorin, C. D. Lisa, M. Ashok Kumar, A. T. Saeed, E. H. Matthew, and M. Seema, "High throughput exome coverage of clinically relevant cardiac genes," *BMC Medical Genomics*, vol. 7, p. 67, 2014. DOI: https://doi.org/10.1186/s12920-014-0067-8. [Online]. Available: https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-014-0067-8.

[9] E. Cornec-Le Gall, A. Alam, and R. D. Perrone, "Autosomal dominant polycystic kidney disease," *New England Journal of Medicine*, vol. 380, no. 2, pp. 171–180, 2019. DOI: 10.1016/S0140-6736(18)32782-X. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S014067361832782X.

[10] C. Bergmann, L. M. Guay-Woodford, P. C. Harris, S. Horie, D. J. Peters, and V. E. Torres, "Polycystic kidney disease," *Nature Reviews Disease Primers*, vol. 4, no. 1, p. 50, 2018. DOI: 10.1038/s41572-018-0047-y. [Online]. Available: https://www.nature.com/articles/s41572-018-0047-y.

[11] G. Negri, D. Milani, P. Colapietro, and et al., "Rubinstein–taybi syndrome: Clinical features, genetic basis, diagnosis, and management," *Italian Journal of Pediatrics*, vol. 45, no. 1, p. 101, 2019. DOI: 10.1186/s13052-015-0110-1. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25599811/.

[12] S. Spena, D. Milani, D. Rusconi, and G. Negri, "Clinical and molecular diagnosis of rubinstein-taybi syndrome," *Expert Review of Molecular Diagnostics*, vol. 15, no. 2, pp. 159–171, 2015. DOI: 10.1111/cge.12348. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24476420/.

[13] J. Kohlhase, A. Wischermann, H. Reichenbach, and et al., "Mutations in the sall1 gene in townes-brocks syndrome," *Nature Genetics*, vol. 18, no. 1, pp. 81–83, 1998. DOI: 10.1038/ng0198-81. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/9425907/.

[14] G. Nalepa and D. W. Clapp, "Fanconi anemia and genomic instability," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 743, pp. 134–140, 2013. DOI: 10.12703/P6-23. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24765528/.

[15] M. C. Kottemann and A. Smogorzewska, "Fanconi anaemia and the repair of watson and crick dna crosslinks," *Nature*, vol. 493, no. 7432, pp. 356–363, 2013. DOI: 10.1038/nature11863. [Online]. Available: https://www.nature.com/articles/nature11863#citeas.

[16] J.-M. Waldburger, K. Masternak, A. Muhlethaler-Mottet, *et al.*, "Lessons from the bare lymphocyte syndrome: Molecular mechanisms regulating mhc class ii expression," *Immunological Reviews*, vol. 178, pp. 148–165, 2000. DOI: 10.1034/j.1600-065X.2000.17813.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1034/j.1600-065X.2000.17813.x?sid=nlm%3Apubmed.