Large pedigrees in human sequencing studies: toward a more resolved and accurate picture of genetic disease

Jason O'Rawe[1], Yiyang Wu[1,2], Alan Rope[3], Laura T. Jimenez Barrón[1,4], Jeffrey Swensen[5], Han Fang[1], David Mittelman[6], Gareth Highnam[6], Reid Robison[7], Kai Wang[7,8], Gholson J. Lyon[1,2,7]

[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, NY, USA; [2]Stony Brook University, Stony Brook, NY, USA; [3]Department of Medical Genetics, Northwest Kaiser Permanente, Portland, OR, USA; [4]Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX; [5]Caris Life Sciences, Phoenix, Arizona, USA; [6]Gene by Gene, Ltd., Hou-ston, TX, USA ; [7]Utah Foundation for Biomedical Research, Salt Lake City, UT, USA, [8]Zilkha Neu-rogenetic Institute, Department of Psychiatry and Preventive Medicine, University of Southern California, Los Angeles, CA, USA

Abstract

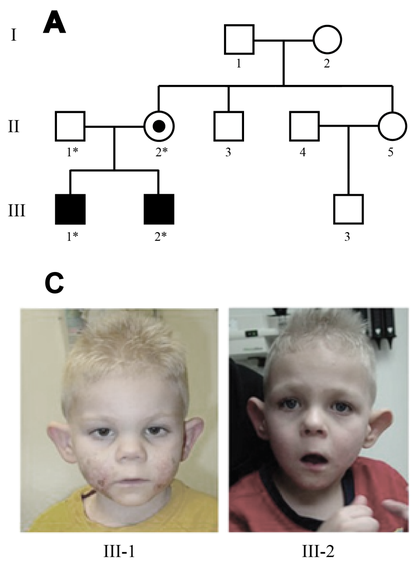
Background: We describe a comprehensive whole genome sequencing (WGS) study using the Illumina and Complete Genomics (CG) sequencing platforms for one family containing two affected male brothers, aged 10 and 12, with severe intellectual disability and very distinctive facial features. High accuracy and sensitivity is of particular importance in the context of detecting or discovering the genetic influencers of human diseases.

Methods: WGS was performed on ten members of this family using the Il-lumina HiSeq2000 platform, with four (the two affected boys and their parents) being additionally sequenced using the CG WGS platform. CG data analysis was performed by CG, using their version 2.0 pipeline. Multiple variant calling pipelines were used to detect SNVs, INDELs, STRs and CNVs. Disease variant prioritization was performed using ANNOVAR, Golden Helix SVS v8.1.4 and GEMINI v0.9.1, and VAAST v2.0.

Results: CG WGS covered >85% of the genome and >90% of the exome, both with 20 or more reads. Illumina WGS covered >90% of the genome with 30 reads or more and with >80% of the bases having a quality score of >30. On average, we find a 2.4 to 14.0 mean fold difference in the number of variants detected as being relevant for various disease models when using different sets of sequencing data and analysis pipelines. We found a number of putative genetic variants and archive them here.

Presentation of the phenotype

- The two affected male brothers have severe intellectual disability, autism-like behavior, attention deficit issues, and very distinctive facial features (**Fig. 1C**), including broad, upturned nose, sagging cheeks, downward sloping palpebral fissures, relative hypertelorism, high-arched palate, and prominent ears.
- The mother of the two affected (**Fig. 1A, II-2**) was shown to have 99:1 x-chromosome inactivation (**Fig. 1B**).



Sample	Allele 1	Allele 2
II-2	0.01	0.99
II-5	0.29	0.71
I-2	0.65	0.35

Fig. 1 (**A**) Pedigree structure of all individuals in the family that were sequenced during the course of this study. Individuals with a star next to their number indicates that their whole genomes were sequenced with both the Complete Genomics sequencing and analysis pipeline as well as with Illumina sequencing.

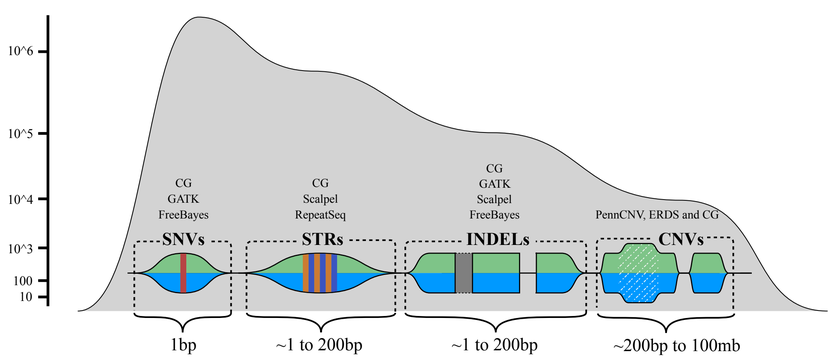
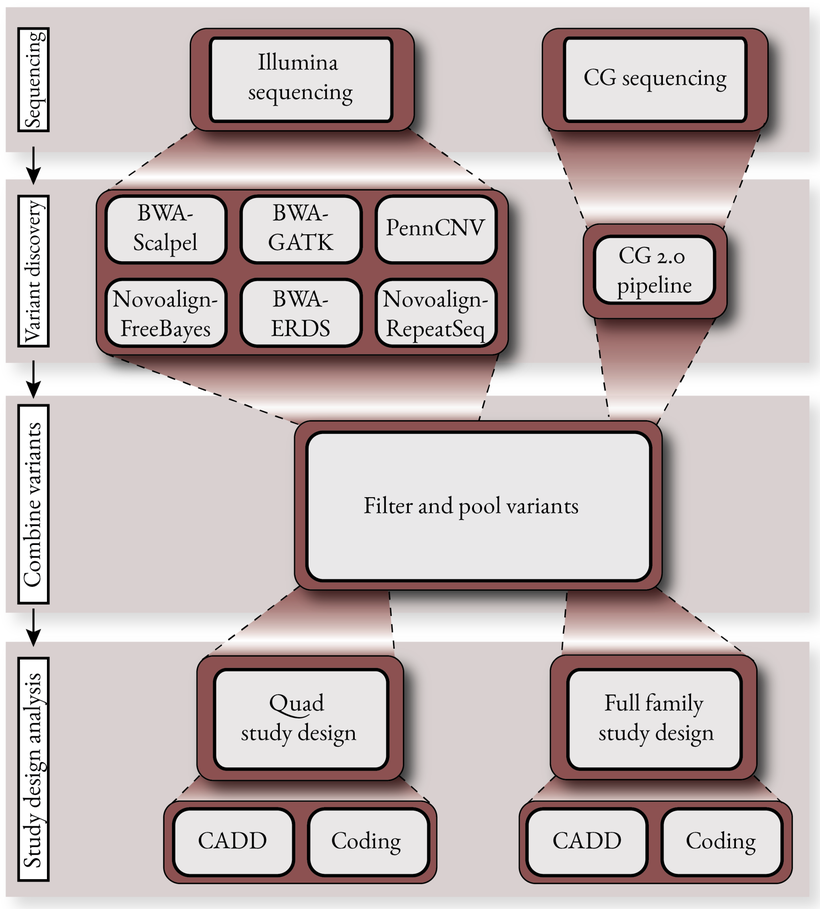
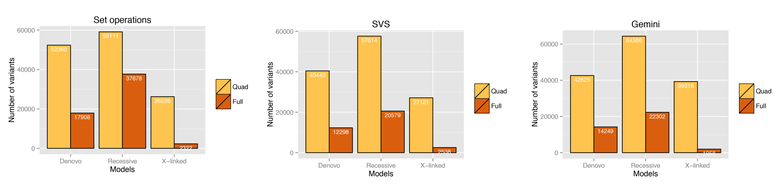


Fig. 2 A conceptual map of human sequence variation, and a list of the bioinformatics programs we used during the course of our study.



Bioinformatics analysis

- Human sequence variation ranges in manifestation from differences that can be detected at the single nucleotide level, to whole chromosome differences (**Fig 2**).
- We sought to identify variants following de-novo, autosomal recessive and x-linked models of transmission that may be contributing, together or alone, to the disease phenotype. We used several methods to prioritize and identify possible disease-contributory germ-line variants, in-cluding VAAST, Golden Helix SVS v8.1.4, ANNOVAR (2013Aug23 version), and GEMINI v0.9.1.



- We found a 2.4 to 14.0 mean fold difference in the number of variants detected as being relevant for vari-ous disease models when using different sets of sequencing data.

Results

- 14 unique INDELS and SNVs were discovered using two different prioritization (**Fig. 3A**).
- The TAF1 variant arose in this family as a de-novo variant on the X-chromosome of the mother (**Fig. 1A, II-2**) of the two affected children (as it is not found in any of the other members of the family) and was then transmitted to both of them. The mother also is the only female in the family to exhibit extreme X-chromosome skewing.
- The transcription factor initiation complex TF11D has recently been implicated in playing a role in intellectual disability and developmental delay, and TAF1 represents the largest known subunit of this multi-protein complex, and this variant falls within a conserved region of the protein (**Fig 3B**).

A

Model	Location	Ref	Alt	Variant Caller	Annotation	Function	Scheme
Recessive	chr1:210851705	TT	T	CG, GATK, FreeBayes, RepeatSeq	ANNOVAR, GEMINI, SVS	KCNH1:UTR3	CADD, score:27.5
Recessive	chr1:224772440	AATAATTTG	TA	CG, GATK, FreeBayes	GEMINI	intergenic	CADD, score:22.1
Recessive	chr2:60537356	TTTTATTT	ATTATTA	CG, FreeBayes, GATK, RepeatSeq	GEMINI	intergenic	CADD, score:22.3
Recessive	chr8:109098066	AT	A	CG, FreeBayes, GATK, RepeatSeq	GEMINI	intergenic	CADD, score:24.6
Recessive	chr15:66786022	ACAAA	A	FreeBayes, GATK	GEMINI	SNAPC5:intronic	CADD, score:23.6
Recessive	chr16:49061346	TA	T	CG, FreeBayes, GATK	ANNOVAR, GEMINI	intergenic	CADD, score:25.3
Recessive	chr16:49612367	GAC	G	CG, FreeBayes, GATK	GEMINI, SVS	ZNF423:intronic	CADD, score:20.5
Recessive	chr10:135438929	T	G	CG, FreeBayes, GATK	ANNOVAR, GEMINI, SVS	NM_001080998: I171L	Coding, gene:FRG2B
Recessive	chr10:135438951	GGCCC	AGCCT	FreeBayes, Scalpel	GEMINI, SVS	NM_001080998: sub	Coding, gene:FRG2B
Recessive	chr10:135438967	C	T	GATK, FreeBayes	GEMINI, SVS	NM_001080998: R158Q	Coding, gene:FRG2B
Recessive	chr15:85438314	C	CTTG	CG, FreeBayes, GATK, Scalpel	GEMINI	NM_201651: K141delinsIE	Coding, gene:SLC28A1
De-novo	chr1:53925373	G	GCCGCCC	FreeBayes, CG, Scalpel	GEMINI, SVS	NM_033067: A83delinsAAP	Coding, gene:DMRTB1
X-linked	chrX:34961492	T	C	CG, FreeBayes, GATK	GEMINI	NM_152631: Y182H	Coding, gene:FAM47B
X-linked	chrX:70621541	T	C	CG, FreeBayes, GATK	ANNOVAR, GEMINI, SVS	NM_004606:I133 7T	Coding, gene:TAF1; CADD, score:22.9

B

```
H.sapiens       1332  D--NEELIKVEGTKIVL--------------------------------  1346
M.mulatta       1244  D--NEELIKVEGTKIVL--------------------------------  1258
C.lupus         1332  D--NEELIKVEGTKIVL--------------------------------  1346
B.taurus        1316  D--NEELIKVEGTKIVL--------------------------------  1330
M.musculus      1343  D--NEELIKVEGTKIVL--------------------------------  1357
R.norvegicus    1332  D--NEELIKVEGTKIVL--------------------------------  1346
G.gallus        1357  D--NEELIKVEGTKIVL--------------------------------  1371
X.tropicalis    1377  D--NEELIKVEGTKIVL--------------------------------  1391
D.rerio         1396  D--D-DLVNVDGTKVTL--------------------------------  1409
D.melanogaster  1419  D--EGDLVNVDGTKVKL--------------------------------  1433
A.gambiae       1263  SLVAPDAVQVDGTKVKFNLNFAEIRKEQNREEKLKRKLAKMAEAAVRERQ  1312
C.elegans       1294  D--NEELIKVEGTKIVL--------------------------------  1308
```

Fig. 3

- *Conclusions:* Analyzing multi-generational pedigrees using multiple orthogonal bioinformatics pipelines using two sequencing platforms can reliably reveal human sequence variants that may be important in rare disease. We have found a number of sequence variants that may play a role in the rare disease described here and highlight a variant in TAF1. Our findings are consistent with the literature on the importance of the TF11D complex in developmental delay and ID.