Clinical whole genome sequencing: a validation study

Gerard Irzyk[1], C. Alexander Valencia[1], Zeqiang Ma[1], Edward Szekeres Jr[1], Zdenek Markovic[1], Dhara Shah[1], Yang Wang[1], Alice Tanner[1], Christin Collins[1], Madhuri Hegde[1,2]
[1]PerkinElmer Genomics, [2]Emory University

ABSTRACT

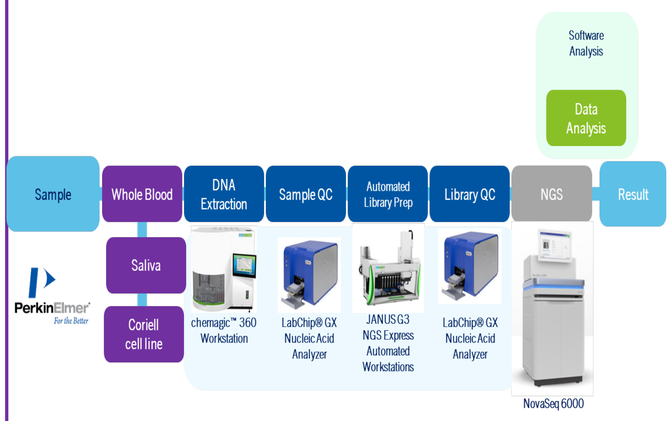
Whole-exome sequencing (WES) is now routinely used in clinical genetic testing and is optimized for the detection of rare and common genetic variants in humans. However, whole genome sequencing (WGS) is becoming increasingly attractive as an alternative, due to its uniform coverage, ability to detect all types of variants and decreasing cost. The objective of this study was to validate WGS in a clinical laboratory using a PCR-free library preparation protocol followed by sequencing on the Illumina NovaSeq™, primary data processing using the Edico Dragen system, and bioinformatic analysis using our in-house proprietary program ODIN. Data was parsed into various categories: Genes causing disease (GCD), Genes of unknown significance (GOUS), ACMG 59, known common and founding pathogenic changes, intragenic and intergenic variants with tagged variants which have been established to be diseases causing. A total of 52 samples, including positive control Coriell samples with known pathogenic variants (26), three control samples from Coriell cell lines NA12878, NA12892 and NA12891 , DNA extracted from whole blood samples (23) obtained from positive controls with known pathogenic variants or DNA extracted from unaffected saliva (3). All samples were sequenced at >30X average coverage and analyzed. Subsequently, performance parameters for accuracy, precision, sensitivity, and specificity were calculated. The average global genome metrics for the 52 samples were 38.56X average coverage and 90.60% coverage of the genome at 10X. In addition, the average number of SNPs and indels detected were 4,132,282 (81.52% of calls) and 936,598 (18.48% of calls), respectively. The average number of homozygous and heterozygous calls were 3,243,145 and 1,825,734, respectively. To calculate inter-run accuracy, NA12878 was run twice and compared to the genome in a bottle (GIAB) NA12878 sample, yielding an accuracy value of 99.7% (3,207,518 SNPs in common). Of the unique SNPs (19,951) found in our NA12878 samples, ~30% were determined to be artifacts by bioinformatics analysis. The precision value of 99.2% was calculated by running 11 samples, including Coriell samples NA12878, NA12892 and NA891, two times. Specifically, there was an average number of 3,379,210 SNPs that were concordant in the duplicate samples. The sensitivity, defined as the percentage of low coverage exons, was 99.00% based on 52 samples. From a total of 63,962 exons (4802 genes associated with disease), an average of only 704 exons had coverage less than 8X belonging to 160 genes, which included *SMN1* and *SMN2* due to high homology. The average specificity was 99.30%. Moreover, in this validation study samples with specific variant(s) were included such as missense, splicing and nonsense variants as well as exon-level deletions and duplications to see if they could be identified. All specific variants were identified by WGS. For example, the expected c.1448T>C (p.L483P) and c.115+1G>A were detected in *GBA*, which has homologous sequence with 99% overlap. Moreover, we detected a hemizygous deletion of exons 8-10 in *ABCD1*, X-linked adrenoleukodystrophy. In this WGS validation study it was demonstrated that all of the performance parameter were ≥99%, thus, it is acceptable as a clinical testing. As we learn more about the non-coding regions of the genome we anticipate better clinical interpretation and thus more powerful re-analysis using WGS data in the near future.

INTRODUCTION

- Whole-exome sequencing (WES) is now routinely used in clinical genetic testing and is optimized for the detection of rare and common genetic variants in humans with an approximately 25-30% diagnostic yield
- **CHALLENGES:** WES can miss major types and regions of disease-causing genomic variation (INDELs, structural variants, intronic SNVs)
- Whole genome sequencing (WGS) is becoming increasingly attractive as a clinical alternative
- **WGS ADVANTAGES:** (i) Higher diagnostic yield than WES, (ii) detection of non-exonic sequence variants, (iii) improved CNV detection and (iv) added lifetime value of re-analysis over time, as clinical features develop in the future of an individual
- **APPROACH:** WGS was performed by using the KAPA HyperPlus PCR-free library construction kit on DNAs obtained from a number of samples types: whole blood, saliva, and Coriell cell lines. The libraries were subsequently sequenced Illumina NovaSeq™ 6000 operating in 2 x 150 bp mode

METHODS

WGS sequencing workflow



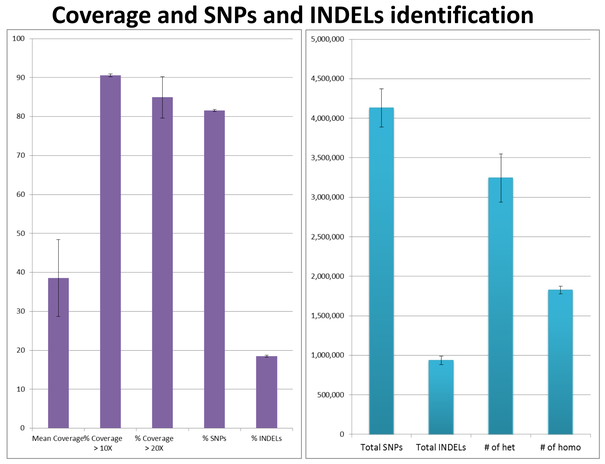
WGS validation plan

Sample Name	Sample type	Accuracy	Precision	Sensitivity	Specificity
NA12878-GIAB	CL	+			
NA12878-1	CL	+	+	+	+
NA12878-2	CL	+	+	+	+
NA12892-1	CL		+	+	+
NA12892-2	CL		+	+	+
NA12891-1	CL		+	+	+
NA12891-2	CL		+	+	+
NA17819-1	CL		+	+	+
NA17819-2	CL		+	+	+
NA03723-1	CL		+	+	+
NA03723-2	CL		+	+	+
NA01361-1	CL		+	+	+
NA01361-2	CL		+	+	+
NA20088-1	CL		+	+	+
NA20088-2	CL		+	+	+
NA04372-1	CL		+	+	+
NA04372-2	CL		+	+	+
43313	WB		+	+	+
43314	WB		+	+	+
5656-1	WB		+	+	+
5656-2	WB		+	+	+
5647-1	WB		+	+	+
5647-2	WB		+	+	+
NA20310	CL			+	+
NA20270	CL			+	+
NA08684	CL			+	+
NA03251	CL			+	+
NA00107	CL			+	+
NA22496	CL			+	+
NA13537	CL			+	+
NA11661	CL			+	+
NA06314	CL			+	+
NA03814	CL			+	+
10	WB			+	+
F8	WB			+	+
7	WB			+	+
5	WB			+	+
5642	WB			+	+
14	WB			+	+
5548	WB			+	+
5415	WB			+	+
3639	WB			+	+
2A	WB			+	+
16	WB			+	+
3069	WB			+	+
5C	WB			+	+
JF216	WB			+	+
B-SLV	SLV			+	+
B-WB	WB			+	+
E-SLV	SLV			+	+
E-WB	WB			+	+
T-SLV	SLV			+	+
T-WB	WB			+	+
Total		3	22	52	52

RESULTS

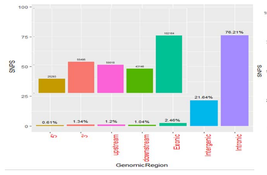
Global Statistics: High quality global key measurements

Coverage and SNPs and INDELs identification

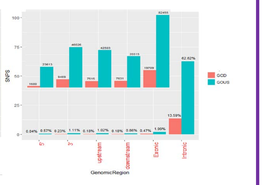


- High coverage and sensitivity detection of SNPs and INDELs

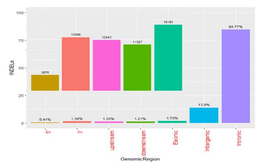
SNPs Distribution over All Genes



SNPs Distribution over GCD/GOUS Genes



INDELs Distribution over All Genes



INDELs Distribution over GCD/GOUS Genes

- Comprehensive detection of SNPs and INDELS across entire genome and genes causing disease (GCD) genes

Excellent performance and comprehensive detection of variants

Accuracy is 99.7%

Terms	# of SNPs defined by GIAB	
	NA12878-1	NA12878-2
False positive (NA12878 only)	19322	20580
False negative (GIAB only)	4076	1388
True positive (Overlap)	3206174	3208862
True negative (Overlap)	3800617	3801969
Accuracy	99.7%	99.7%
Average accuracy	99.7%	

*GIAB – not confirmed by orthogonal method

Sensitivity is 99.0%

Sample Name	Genes	Transcripts	% Exons	Exons Coverage Genes	Exons Coverage Transcripts	% Low Coverage Transcripts	% Low Coverage Exons	% Sensitivity
NA12878-1	4802	4802	63962	103	2.1	103	2.1	577	0.9	99.1
NA12878-2	4802	4802	63962	65	1.4	65	1.4	494	0.8	99.2
NA12892-1	4802	4802	63962	67	1.4	67	1.4	508	0.8	99.2
NA12892-2	4802	4802	63962	93	1.9	93	1.9	558	0.9	99.1
NA12891-1	4802	4802	63962	76	1.6	76	1.6	438	0.7	99.3
NA12891_2	4802	4802	63962	208	6.0	208	6.0	1086	1.7	98.3
NA17819-1	4802	4802	63962	69	1.4	69	1.4	420	0.7	99.3
NA17819-2	4802	4802	63962	299	6.2	299	6.2	1069	1.7	98.3
NA03723-1	4802	4802	63962	70	1.5	70	1.5	536	0.8	99.2
NA03723-2	4802	4802	63962	105	2.2	105	2.2	596	0.9	99.1
NA01361-1	4802	4802	63962	63	1.3	63	1.3	411	0.6	99.4
NA01361-2	4802	4802	63962	280	5.8	280	5.8	1051	1.6	98.4
NA20088-1	4802	4802	63962	78	1.6	78	1.6	536	0.8	99.2
NA20088-2	4802	4802	63962	97	2.0	97	2.0	573	0.9	99.1
NA04372-1	4802	4802	63962	410	8.5	410	8.5	1693	2.6	97.4
NA04372-2	4802	4802	63962	231	4.8	231	4.8	802	1.3	98.7
B-1	4802	4802	63962	140	2.9	140	2.9	623	1.0	99.0
B-2	4802	4802	63962	121	2.5	121	2.5	603	0.9	99.1
Average	4802	4802	63962	165	3.3	165	3.3	704	1.1	98.9

Genomes from 52 samples provided average of 161 (3.3%) genes or 704 exons (1.1%) with coverage of <8X

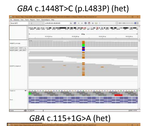
Precision is 99.2%

Sample Name	Precision (using all SNPs from genome) # of SNPs			Precision (using GIAB SNPs from genome) Percentage			# of SNPs			Percentage		
	Unique Sample 1	Unique Sample 2	overlap	Unique Sample 1	Unique Sample 2	Overlap (precision)	Sample 1	Sample 2	overlap	Sample 1	Sample 2	overlap (precision)
NA17819	116291	59624	3697321	2.8	1.7	95.4	16479	9309	3122441	0.5	0.3	99.2
NA12892	106217	78861	3921387	2.6	1.9	95.5	12883	8444	3246818	0.4	0.3	99.3
NA12891	111370	69835	3891132	2.7	1.7	95.6	13180	7694	3198251	0.4	0.2	99.4
NA12878	72320	115627	3926174	1.8	2.8	95.4	7529	11475	3217767	0.2	0.4	99.4
NA04372	92291	112962	4566306	1.9	2.4	95.7	13129	18139	3669242	0.3	0.5	99.2
NA03723	110576	76738	4646838	2.4	1.6	96.0	16365	9159	3915366	0.4	0.2	99.3
NA01361	141852	78737	3933357	3.4	1.9	94.6	36385	19820	3164545	0.9	0.5	98.3
5656	123795	71498	4003268	2.9	1.7	95.3	16122	8781	3319445	0.5	0.3	99.3
5647	117436	74455	4042741	2.8	1.8	95.5	16202	9038	3360388	0.5	0.3	99.3
4	94321	96126	3891339	2.3	2.4	95.3	12496	13151	3239598	0.4	0.4	99.2
NA20088	119501	86469	4016329	2.8	2.0	95.1	15206	10543	3342656	0.5	0.3	99.2
Average	**110379**	**84851.09**	**4002306**	**2.6**	**2.0**	**95.4**	**15060**	**11399**	**3379211**	**0.5**	**0.3**	**99.2**

Specificity is 99.3%

Sample Name	# Target Bases	# Fully Covered Target Bases	% Fully Covered Target Bases (Specificity)
NA12878-1	12047455	11954036	99.2
NA12878-2	12047455	11962654	99.3
NA12891-1	12047455	11975987	99.4
NA12891-2	12047455	11946841	99.2
NA12892-1	12047455	11960770	99.3
NA12892-2	12047455	11955766	99.2
NA17819-1	12047455	11977365	99.4
NA17819-2	12047455	11945244	99.2
NA03723-1	12047455	11957922	99.3
NA03723-2	12047455	11950614	99.2
NA01361-1	12047455	11979185	99.4
NA01361-2	12047455	11947144	99.2
NA20088-1	12047455	11956684	99.2
NA20088-2	12047455	11954220	99.2
NA04372-1	12047455	11910601	98.9
NA04372-2	12047455	11961614	99.3
B-1	12047455	11950817	99.2
B-2	12047455	11951657	99.2
5656-1	12047455	11974495	99.4
5656-2	12047455	11919623	98.9
Average	**12047455**	**11957679**	**99.3**

GBA c.1448T>C (p.L483P) (het)



GBA c.115+1G>A (het)

SMN1 suggestive of deletion (het)

Detection of a spectrum of variants

Sample Name	Disorder (Inheritance)	Gene	Alleles	
NA17819	Adrenoleukodystrophy (AR)	*ABCD1*	Exons 8-10 deletion (Homo)	Deletion
NA03723	Mucopolysaccharidosis type VI (Maroteaux-Lamy) (AR)	*ARSB*	c.1433delA (p.D478fs) (Homo)	
NA01361	Mucopolysaccharidosis IVA (AR)	*GALNS*	c.181C>T (p.R61W) (Het) c.1213_1218delTGGACC (p.W405_T406del) (Het)	
NA20088	Glycogen storage disease II (AR)	*GAA*	IVS18 + 2T>A (not detected) 8.26 kb deletion extending from IVS7 to IVS15 (Homo)	
NA04372	Krabbe disease (AR)	*GALC*	30kb deletion (Het)	
8	Krabbe disease (AR)	*GALC*	Exons 11-17 deletion (Homo) common 30 kb deletion (Het)	
5656	Becker muscular dystrophy/ Duchenne muscular dystrophy (XLR)	*DMD*	Exons 45-55 deletion (Hemi)	
5647	Becker muscular dystrophy/ Duchenne muscular dystrophy (XLR)	*DMD*	Exon 18 duplication	Duplication
NA20310	Glycogen storage disease II (AR)	*GAA*	9-kb deletion extending from IVS15 to 4kb downstream of exon 20 (Homo)	
NA20270	Gaucher disease (AR)	*GBA*	c.1448T>C (p.L483P) (Het) c.115+1G>A (Het)	Missense/ Pseudogene
NA08684	Acyl-CoA dehydrogenase, medium chain, deficiency of (AR)	*ACADM*	c.446G>A (p.C149Y) (Het) c.1084A>G p.K362E (Het)	
NA03251	GM1-ganglioisidosis I-III; Mucopolysaccharidosis type IVB (Morquio)	*GLB1*	c.1527G>T (p.W509C) (Het) c.818G>T (p.W273L) (Het)	
NA00107	Fabry disease (XL)	*GLA*	c.485G>A (p.Trp162*) (Hemi) c.640-16A>G (Hemi) c.1000-22C>T (Hemi)	Nonsense
NA22496	Propionicacidemia (AR)	*PCCB*	c.1218_1231delinsTAGAGCACAGGA (c.1218del14ins12) (Homo)	
NA13537	Spinocerebellar ataxia 1 (AD)	*SCA1*	60 CAG repeats 32 CAG repeats	TNR
NA11661	Glycogen storage disease II (AR)	*GAA*	IVS1-13T>G (Het) Exon18 deletion	
NA06314	Glycogen storage disease II (AR)	*GAA*	IVS1-13T>G (Het) c.1551+1G>C (Het)	Splice site
NA03814	Spinal muscular atrophy-1	*SMN1*	Exons 7-9 deletion (Het)	

DISCUSSION/CONCLUSIONS

- NGS in the past decade has led to an enormous increase in the understanding of the human genome and its relation to disease
- Improved technologies continuously provide faster, cheaper and more accurate results, allowing us to move from gene panels to WES to routinely sequencing WGS in the clinic
- In this WGS validation study it was demonstrated that all of the performance parameters were **≥99%, thus, it is acceptable as a clinical test**
- We have successfully used the documented clinical cases to demonstrate that WGS performance translates to a useful clinical sensitivity
- We demonstrated the analytical sensitivity for SNVs and small indels and demonstrated that intra- and inter-gene CNVs were accurately detected by improved calling algorithms, thus, increasing the diagnostic yield of genetic disease burden
- As we learn more about the non-coding regions of the genome, large scale efforts, for example gnomAD, we anticipate improved clinical interpretation and thus more powerful re-analysis using WGS data.