

Carrier Screening of Recessive Genetic Disorders by Next Generation Sequencing

R.W. Kim (rwk@ncgr.org), D.L. Dinwiddie, C.J. Bell, N.A. Miller, B.J. Rice, J.A. Crow, E.E. Ganusova, S.L. Hateley, S.F. Kingsmore

National Center for Genome Resources, Santa Fe, New Mexico



Abstract

Human recessive genetic diseases are individually quite rare, but together are a major medical burden that cause significant morbidity and mortality. 20-30% of all infant deaths and 11% of pediatric hospital admissions are related to genetic disorders. In collaboration with the Beyond Batten Disease Foundation, NCGR is developing a carrier screening test for 448 autosomal recessive (AR) and X-linked recessive (XLR) disorders caused by tens of thousands of mutations. The screening test will utilize target enrichment of genes, multiplexed deep sequencing, and automated bioinformatic analysis to identify carrier status for the 448 selected disorders. To identify the appropriate target enrichment technology for the test, a total of 24 samples from carriers of 17 autosomal recessive diseases and 1 X-linked recessive disease were enriched for 437 genes using Agilent SureSelect™ and RainDance RainStorm™ enrichment techniques and subjected to multiplexed deep sequencing using the Illumina® GA Ix. The 24 samples contained 42 previously characterized mutations including 18 SNPs, 12 short indels, and 7 gross deletions impacting both coding and splicing elements. The two enrichment techniques were evaluated for percent of sequencing reads on target, fold enrichment, mutation detection, required sequencing to achieve 20x coverage at >99% of targeted regions, and sensitivity and specificity of variant detection. Additionally, Illumina® GA Ix and Life Technologies SOLiD™ 3 sequencing were evaluated for heterozygosity detection in 6 samples enriched for target loci with Agilent SureSelect™. The carrier screening test will be available in early 2010 and cost less than \$1,000.

Introduction

- Rare or orphan diseases (prevalence <1/1500 individuals) are estimated to account for 20-30% of all infant deaths and 11% of pediatric hospital admissions.
- Online Mendelian Inheritance in Man (OMIM) lists 4,250 genetic disorders and 2,045 suspected genetic disorders.
- While tests are available for individual orphan diseases, development of a carrier screen for hundreds of disorders simultaneously would be extremely beneficial.
- We are currently developing a carrier screening test that can simultaneously screen for mutations that cause 448 XLR and AR disorders.

Methods

- 7,717 genomic regions comprising 1,978,041 bp were submitted to Agilent eArray and RainDance for bait and primer design, respectively.
- Genomic DNA from carriers of 17 autosomal recessive diseases and 1 X-linked recessive disease was obtained from Coriell Cell Repository and was enriched for 437 genes using custom designed Agilent SureSelect™ and RainDance RainStorm™ assays.
- Samples were 12-plexed and sequenced on 7 lanes of Illumina GA Ix flowcell with a read length of 50 bp.
- Coverage data, enrichment stats, and detection of known mutations were completed with the Alpheus® Sequence Variant Detection pipeline for all 24 samples.

Methods (Continued)

- To compare sequencing platforms, 6 additional genomic samples from Coriell were enriched with the custom designed Agilent SureSelect kit and sequenced with Illumina® GA Ix and Life Technologies SOLiD™ 3 platforms.
- All SNPs in the targeted region were identified. If the number of reads calling a SNP was between 15-85% of the total number of reads the variant was described as heterozygous.

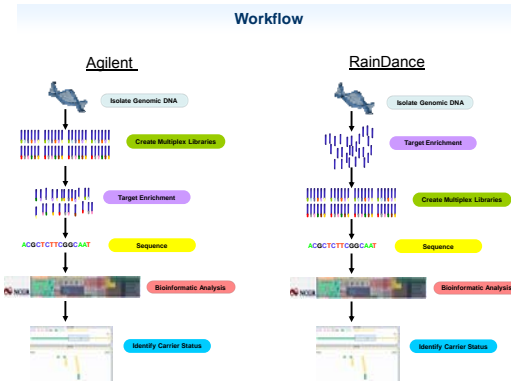


Figure 1. Comparison of the workflow for Agilent SureSelect™ and RainDance Rainstorm™ multiplex enrichment and sequencing.

Results

In Silico Design

	Regions Submitted	Nucleotides Submitted	Regions Designed For	% successful	Notes
Agilent	7,717	1,978,041	7,616	98.69	29,262 primer pairs, 9,824 amplicons
RainDance	7,717	1,978,041	7,646	99.08	5,337,010 total amplicon bases

Table 1. Summary of In Silico design results for 7,717 genomic regions (1,978,041 bp) of target submitted.

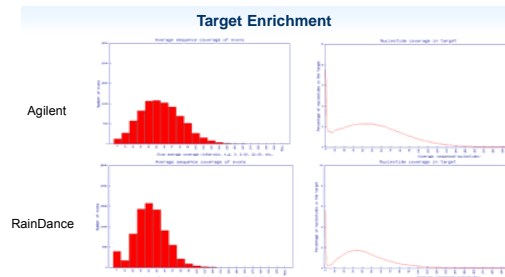


Figure 2. Average coverage depth of exons. Number of reads covering each nucleotide from target (representative graphs from one sample).

Results (Continued)

	Total reads	Aligning reads	Uniquely aligning reads	% unique reads	Aligning bp	Total depth	bp on target	% bp on target	Enrichment	C0	C20	C100	Median coverage	Mean coverage	Coverage SD
RainDance															
Mean	9,967,275	8,155,447	7,899,907	97	407,772,371	252	121,376,163	30	462	5	86	16	56	61	73
Median	9,412,698	7,741,072	7,489,846	97	387,053,600	238	114,052,401	30	462	5	86	13	53	58	69
SD	2,842,253	2,397,275	2,327,289	0	115,863,740	74	39,104,317	2	25	0	4	13	18	20	23
CV	30	29	29	0	29	30	32	5	5	2	5	87	33	32	32
Agilent															
Mean	9,988,157	9,135,923	8,690,323	95	456,796,129	253	105,322,005	23	360	4	79	11	49	53	64
Median	10,127,721	9,231,433	8,792,217	95	461,574,650	256	106,917,486	23	358	4	80	12	50	54	66
SD	1,583,108	1,458,064	1,385,795	0	72,753,197	40	16,372,635	0	7	0	4	6	8	8	10
CV	16	16	16	0	16	16	2	2	7	5	52	16	16	15	15

Table 2. Enrichment statistics of 24 samples enriched with RainDance RainStorm™ or Agilent SureSelect™ and sequenced using an Illumina GA Ix.

Heterozygous Variant Detection

Sample ID	Mutation	Type	Chr	Start	Stop	Allele	Gene	Disease	RainDance	% SureSelect	%	
NA1868	CD800142	deletion	7	116980881	116988884		CFTR	Cystic fibrosis	20/33	61%	29/52	56%
NA1869	CD800142	deletion	7	116980881	116988884		CFTR	Cystic fibrosis	9/24	38%	14/28	57%
NA20379	CM81625	substitution	1	40330765	40330766	T	PF11	Neuronal Ceroid Lipofuscinosis 1	21/42	50%	21/48	44%
NA20379	CM950975	substitution	1	40332957	4032957	A	PF11	Neuronal Ceroid Lipofuscinosis 1	13/38	49%	49/95	52%
NA20382	CD721420	deletion	16	28406313	28406314	CTA	CLN3	Neuronal Ceroid Lipofuscinosis 3	22/50	44%	8/19	42%
NA20383	CM003663	substitution	16	28401322	28401322	A	CLN3	Neuronal Ceroid Lipofuscinosis 3	64/119	54%	11/21	52%
NA20375	CM900051	substitution	15	70429913	70429913	T	HEXA	Tay Sachs disease	33/75	44%	15/26	58%
NA21496	CM90055	insertion	11	520440	520440	T	HEB	Thalassemia beta	28/57	49%		
NA11110	CM910234	substitution	12	10175832	10175832	C	PAH	Phenylketonuria	38/81	47%	19/50	38%
NA11277	CM900275	deletion	7	116988879	116988882		CFTR	Cystic fibrosis	14/35	40%	1/18	17%
NA13591	CM890142	deletion	7	116988881	116988884		CFTR	Cystic fibrosis	24/52	46%	14/26	54%
NA13591	CM900043	substitution	7	116985262	116985265	A	CFTR	Cystic fibrosis	16/53	30%	46/81	57%
NA13591	CM952819	substitution	1	11778965	11778965	A	MTHFR	Homocystinuria	62/100	62%	27/54	50%
NA18681	CM81629	substitution	1	40327754	40327754	A	PF11	Neuronal Ceroid Lipofuscinosis 1	20/35	57%	36/70	51%
NA18681	CM91827	substitution	1	40330430	40330430	C	PF11	Neuronal Ceroid Lipofuscinosis 1	6/20	30%	24/58	41%
NA18193	CM910355	substitution	11	6372010	6372010	T	SMPD1	Niemann-Pick disease	55/104	53%	21/61	34%
NA18193	CD10554	deletion	11	6372345	6372348	SMPTD	SMPD1	Niemann-Pick disease	36/51	71%	17/36	47%
NA18193	CM880036	substitution	1	15347258	15347258	C	GBA	Gaucher disease 1	19/63	30%	31/114	27%
NA20049	CM900522	substitution	19	4662237	4662237	A	BCDH4	Majeed syndrome	42/95	49%	15/36	42%
NA20049	CD91612	deletion	19	4620380	4620388	CCDHA	CCDHA	Majeed syndrome	52/60	37%	23/46	50%
NA20395	CM940801	substitution	17	75701316	75701316	A	GAA	Glycogen storage disease 2	15/38	39%	6/13	55%
NA20213	CD06064	deletion	12	100671378	100671380	GNPTAB	GNPTAB	Mucopolysac II	44/113	39%	36/98	37%
NA20275	CM92421	insertion	1	23406022	23406023	C	LYST	Chediak-Higashi syndrome	24/42	57%	16/54	30%
NA20275	CM870031	substitution	20	42682446	42682446	A	ADA	Adenosine deaminase deficiency	21/58	36%	21/39	54%
NA20395	CM930285	substitution	17	75706665	75706665	T	GAA	Glycogen storage disease 2	46/92	50%	17/37	45%

Table 3. Number and percent of reads from Alpheus® Sequence Variant Detection pipeline that have variant or wild type sequence at known heterozygous variant position.

Homozygous Variant Detection

Sample ID	Mutation	Type	Chr	Start	Stop	Allele	Gene	Disease	RainDance	% SureSelect	%	
NA13591	CM90827	substitution	6	26199158	26199158	G	HEF	Haemochromatosis	148/148	100%	95/96	99%
NA20257	CM910011	substitution	4	17859912	17859912	E	ASGA	Aspartylglucosaminuria	14/15	93%	6/6	100%
NA20257	CM910010	substitution	4	17859918	17859918	T	ASGA	Aspartylglucosaminuria	14/14	100%	6/6	100%

Table 4. Number and percent of reads from Alpheus® Sequence Variant Detection pipeline that have variant or wild type sequence at known homozygous variant position.

Illumina vs SOLiD Sequencing of SureSelect Enriched Samples

	Total reads	Aligning reads	Uniquely aligning reads	% unique reads	Total	bp on target	% bp on target	Enrichment	C0	C20	C100	Median coverage	Mean coverage	Coverage SD
Illumina														
Mean	21,028,828	18,395,241	18,469,440	95	969,762,050	532	167,116,299	17	271	2	85	34	79	85
Median	19,711,735	18,196,922	17,316,639	95	905,845,375	499	158,207,553	17	271	2	86	34	76	80
SD	7,719,927	6,515,714	6,197,402	0	325,785,707	180	50,917,093	1	8	0	4	16	24	26
CV	34	34	34	0	34	34	3	3	16	4	47	31	30	30
SOLiD														
Mean	9,925,333	9,925,333	8,261,210	83	496,266,658	251	103,980,819	21	327	6	73	13	47	53
Median	9,962,035	9,962,035	8,321,032	83	497,601,725	252	105,042,528	21	330	5	73	13	48	53
SD	531,615	531,615	436,331	0	26,580,747	13	6,334,123	1	18	0	1	3	3	3
CV	5	5	5	0	5	5	6	6	6	2	22	6	6	6

Table 5. Enrichment statistics of 6 samples enriched with Agilent SureSelect custom designed kit and sequenced 6-plex using an Illumina GA Ix and Life Technologies SOLiD™ 3.

Results (Continued)

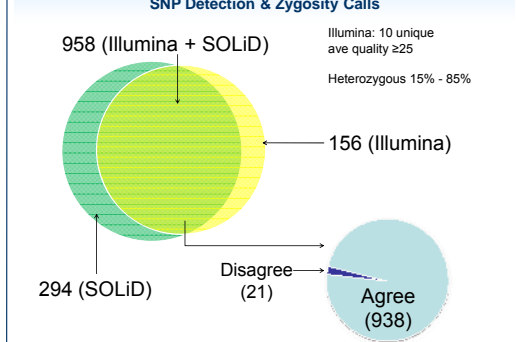


Figure 3. SNP detection and zygosity calls from Illumina and SOLiD sequencing.

Summary

- Both Agilent eArray and RainDance bioinformatic tools were able to design baits or primers for ~99% of targeted regions.
- An average of 30% & 23% of sequenced base pairs were on the target region for RainDance & Agilent resulting in an average fold enrichment of 462 and 360, respectively.
- The Alpheus® Sequence Variant Detection pipeline was able to correctly identify heterozygous and homozygous SNPs & in/dels and call carrier status.
- Illumina and SOLiD™ sequencing of SureSelect™ enriched samples had high concordance in variant and zygosity calls.

Conclusions

Carrier status of rare recessive disorders can be accurately identified by target enrichment of approximately 2Mb of genomic loci with either Agilent SureSelect or RainDance RainStorm™ technologies followed by Illumina GA Ix or SOLiD™ 3 next-generation sequencing using the Alpheus® Sequence Variant Detection pipeline.

Acknowledgements

We thank BBDF for support, Emily Leproust (Agilent), Clarence Lee, Jessica Spangler, Christopher Clousner, Vrunda Sheth, Heather Peckham (Life Technologies), Lin Pham, Keith Brown, James Brayer, Take Ogawa, Steven Kotsopoulos (RainDance), Jimmy Woodward, Peter Ngam, Jenny van Velkinburgh, Ray Langley, Forrest Black, Kathy Myers, John Utsey, and Greg May (NCGR) for their help.