Targeted sequencing of highly homologous genes using Cas9 digestion

Chris Boles, Christine Vasselin, Ezra Abrams

Sage Science, Inc., Beverly, MA 01915

Abstract

Next generation sequencing (NGS) is now widely used for genetic testing in clinical settings. Genes with high homology to other regions of the genome present major challenges for current short-read targeted sequencing methods. We have developed a new method for addressing this problem, which relies on the programmable, high-specificity DNA cleavage of the Cas9 enzyme. The process involves four steps: a) isolation of extremely high molecular weight (HMW) DNA, b) digestion of the HMW DNA with customized Cas9 enzymes that cleave unique sites at the boundaries of the DNA sequencing target, c) size selection electrophoresis to isolate the sequencing target, and d) sequencing of the target using an NGS instrument. To enable this method in clinical settings, we have developed an instrument that can accept isolated white blood cells as the input sample and perform DNA extraction, Cas9 digestion, and preparative target purification in an integrated, semi-automated fashion. We report NGS sequencing results on a medically important gene, PKD1, that has six closely related pseudogenes with sequence homology to about ¾ of the functional gene copy.

Medically relevant genes in regions of high homology

NGS dead zone

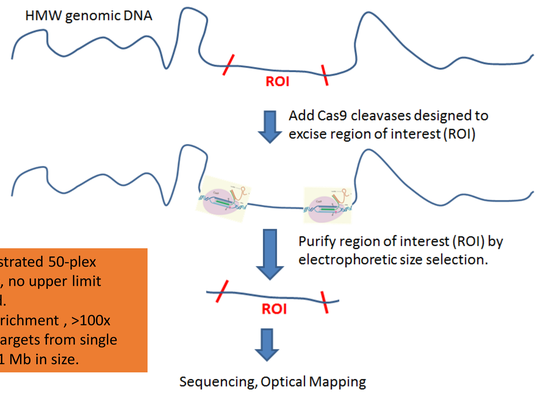
Gene	Affected exons (%)	Affected positions (%)	% Observed low MQ	Homology type	Disease(s)	Category
SMN1	14/16 (87.5)	3,488/3,800 (90.6)	92.7	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
RPS17	8/10 (80)	1,850/2,116 (87.4)	76.4	Same gene | non-CDS | pseudogene	Diamond-blackfan anemia	M
SMN2	14/18 (77.8)	3,488/4,140 (84.3)	93.0	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
IKBKG	7/10 (70)	1,921/2,764 (69.5)	63.6	Pseudogene	Incontinentia pigmenti	M
CFC1	5/6 (83.3)	837/1,471 (56.9)	76.7	Different gene	Congenital heart defects	M
ADAMTSL2	9/18 (50)	2,738/5,196 (52.7)	60.0	Non-CDS	Geleophysic dysplasia	M
OPN1MW	7/12 (58.3)	1,915/3,750 (51.1)	67.2	Same gene | different gene	Colorblindness, deutan; blue cone monochromacy	M
STRC	10/29 (34.5)	3,987/9,098 (43.8)	80.9	Different gene | non-CDS | pseudogene	Sensorineural hearing loss	M
KRT86	3/9 (33.3)	904/2,631 (34.4)	37.9	Different gene | non-CDS | pseudogene	Monilethrix	M
TUBB2B	1/4 (25)	528/1,858 (28.4)	72.0	Different gene | non-CDS | pseudogene	Polymicrogyria	M
LPA	10/39 (25.6)	3,003/11,193 (26.8)	39.5	Same gene | different gene | non-CDS | pseudogene	Coronary artery disease	A
CHRNA7	2/10 (20)	668/2,896 (23.1)	59.6	Different gene	15q13.3 microdeletion syndrome	M
KRT81	2/9 (22.2)	474/2,688 (17.6)	36.0	Different gene | non-CDS | pseudogene	Monilethrix	M
NCF1	2/11 (18.2)	454/2,603 (17.4)	22.3	Pseudogene	Chronic granulomatous disease	M
OTOA	4/30 (13.3)	1,124/7,358 (15.3)	28.5	Non-CDS	Sensorineural hearing loss	M
KIR3DL1	1/9 (11.1)	346/2,505 (13.8)	41.9	Different gene | non-CDS | pseudogene	HIV disease progression	A
TNXB	10/56 (17.9)	2,890/21,942 (13.2)	25.5	Same gene	Ehlers-danlos syndrome	M
OPN1LW	1/6 (16.7)	241/1,875 (12.9)	10.8	Different gene	Blue cone monochromacy	M
NEB	16/181 (8.8)	4,786/49,213 (9.7)	15.3	Same gene	Nemaline myopathy	M
CORO1A	1/10 (10)	235/2,686 (8.7)	7.9	Non-CDS	Immunodeficiency	M
OCLN	1/8 (12.5)	172/2,609 (6.6)	36.0	Pseudogene	Band-like calcification with simplified gyration and polymicrogyria	M
FLG	1/2 (50)	802/12,446 (6.4)	20.0	Same gene	Ichthyosis vulgaris	M
HYDIN	6/86 (7)	1,701/26,643 (6.4)	67.4	Non-CDS | pseudogene	Primary ciliary dyskinesia	M
RHCE	1/10 (10)	157/2,554 (6.1)	17.4	Different gene	Rh blood group antigens	M
PMS2	1/15 (6.7)	274/4,539 (6)	20.4	Non-CDS | pseudogene	HNPCC	M, S
STAT5B	1/18 (5.6)	266/4,704 (5.7)	15.0	Different gene	Growth hormone insensitivity with immunodeficiency	M
TTN	7/363 (1.9)	1,308/161,621 (0.8)	2.2	Same gene	Dilated cardiomyopathy	M

Sanger dead zone

Gene	Affected exons (%)	Affected positions (%)	% Observed low MQ	Homology type	Disease(s)	Category
RPS17	6/10 (60)	1,286/2,116 (60.8)	76.4	Same gene | non-CDS | pseudogene	Diamond-blackfan anemia	M
SMN1	12/16 (75)	2,310/3,850 (60)	92.7	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
SMN2	12/18 (66.7)	2,310/4,140 (55.8)	93.0	Different gene | non-CDS | pseudogene	Spinal muscular atrophy	M
ADAMTSL2	9/18 (50)	2,736/5,196 (52.7)	60.0	Non-CDS	Geleophysic dysplasia	M
IKBKG	6/10 (60)	1,447/2,764 (52.4)	63.6	Pseudogene	Incontinentia pigmenti	M
OPN1MW	5/12 (41.7)	1,431/3,750 (38.2)	67.2	Same gene | different gene	Colorblindness, deutan; blue cone monochromacy	M
CFC1	3/6 (50)	387/1,471 (26.3)	76.7	Different gene	Congenital heart defects	M
OPN1LW	2/6 (33.3)	346/1,875 (18.5)	10.8	Different gene	Blue cone monochromacy	M
STRC	4/29 (13.8)	1,484/9,098 (16.3)	80.9	Different gene | non-CDS | pseudogene	Sensorineural gearing loss	M
CORO1A	2/10 (20)	344/2,686 (12.8)	7.9	Non-CDS	Immunodeficiency	M
GRK1	2/7 (28.6)	300/2,602 (11.5)	0.0	Different gene	Oguchi disease	M
LPA	6/39 (15.4)	1,220/11,193 (10.9)	39.5	Same gene | different gene | non-CDS | pseudogene	Coronary artery disease	A
OCLN	2/8 (25)	213/2,609 (8.2)	36.0	Pseudogene	Band-like calcification with simplified gyration and polymicrogyria	M
NCF1	1/11 (9.1)	202/2,603 (7.8)	22.3	Pseudogene	Chronic granulomatous disease	M
RHCE	1/10 (10)	157/2,554 (6.1)	17.4	Different gene	Rh blood group antigens	M
NEB	11/181 (6.1)	2,560/49,213 (5.2)	15.3	Same gene	Nemaline myopathy	M
OTOA	2/30 (6.7)	375/7,358 (5.1)	28.5	Non-CDS	Sensorineural hearing loss	M
TNXB	4/56 (7.1)	406/21,942 (1.9)	25.5	Same gene	Ehlers-danlos syndrome	M
PMS2	1/15 (6.7)	67/4,539 (1.5)	20.4	Non-CDS | pseudogene	HNPCC	M, S

Genes for which design of **hybridization baits** (NGS dead zone) or **PCR amplicons** (Sanger dead zone) are complicated by segmental duplication or pseudogenes.
(From Mandelker ,et al., 2016, Genetics In Medicine 18(2): 1282-1289.)

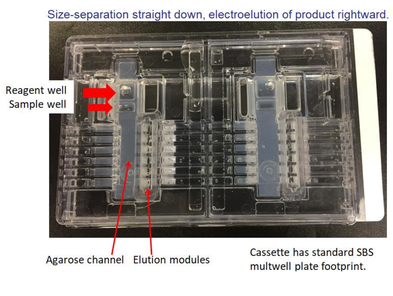
CATCH concept for isolation of genomic DNA targets



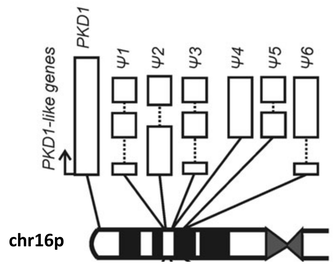
- Demonstrated 50-plex capture, no upper limit reached.
- High enrichment , >100x
- Enrich targets from single Kb's to 1 Mb in size.

(Jiang et al., Nature Communications, 2015, 6:8101)

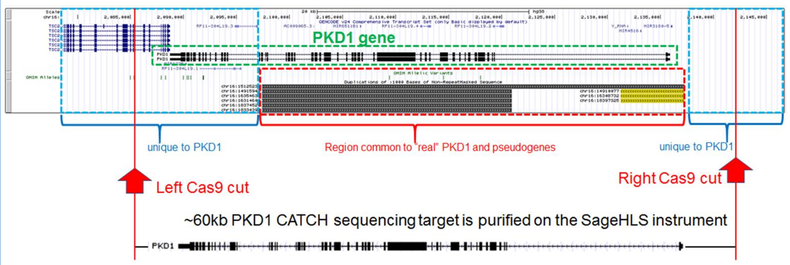
SageHLS Cassette and Instrument



Cas9 guide RNA design for PKD1, a gene with six pseudogenes



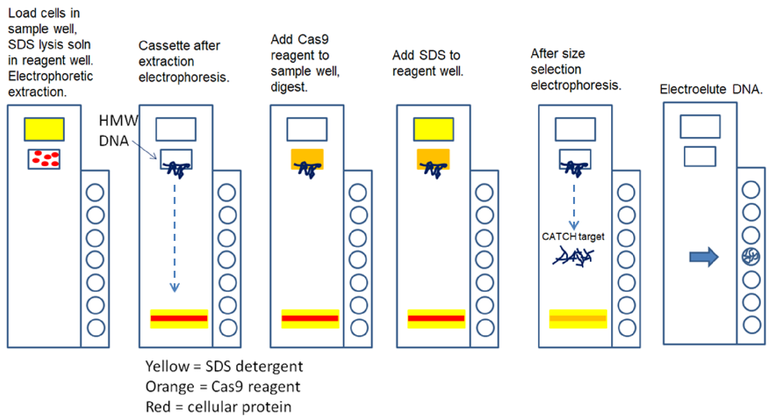
From Symmons et al., 2008 Mol.Biol.Evol.25(12:2601-2613.



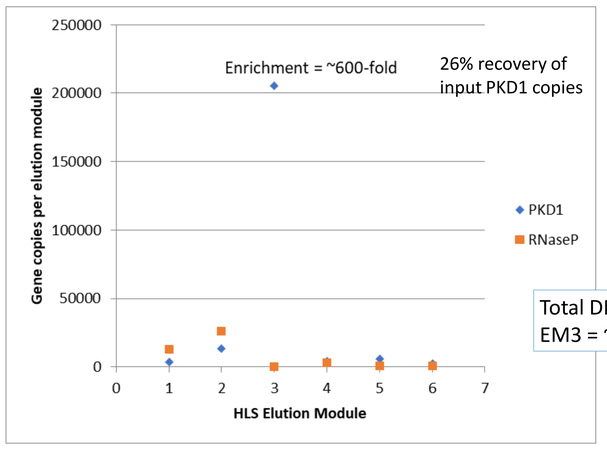
About ¾ of the active PKD1 gene sequence is shared with six pseudogenes, all located on chr16. To excise the active gene , we select gRNA sequences from unique sequence regions that flank the regions that are duplicated in the various PKD-pseudogenes. Using this strategy, a ~60kb CATCH target can be isolated and sequenced without complications from the six other repeated regions.

Integrated DNA extraction, digestion, and size-selection using HLS-CATCH

~375,000 NA12878 human lymphoblastoid cells were loaded into the sample well of an HLS cassette. A lysis buffer containing SDS was loaded into the upstream reagent well. Electrophoretic extraction was carried out for 3 hours, so that the SDS was driven through the sample well. After extraction, the HMW DNA, which remained trapped in the sample wall wall, was digested with Cas9-gRNA complexes specific for the PKD1 gene. After digestion, the excised PKD1 gene was recovered using an automated HLS size-selection + electroelution program.



Enrichment of the PKD1 CATCH product as measured by qPCR

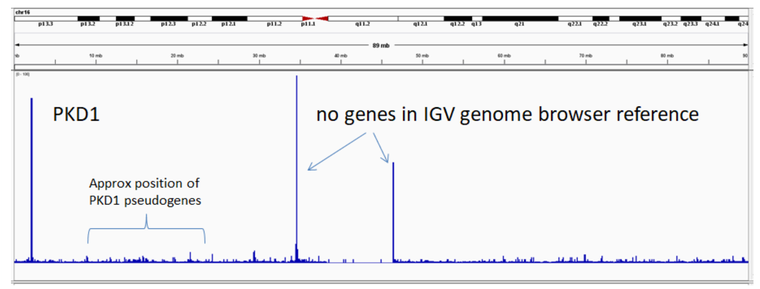


Sequencing

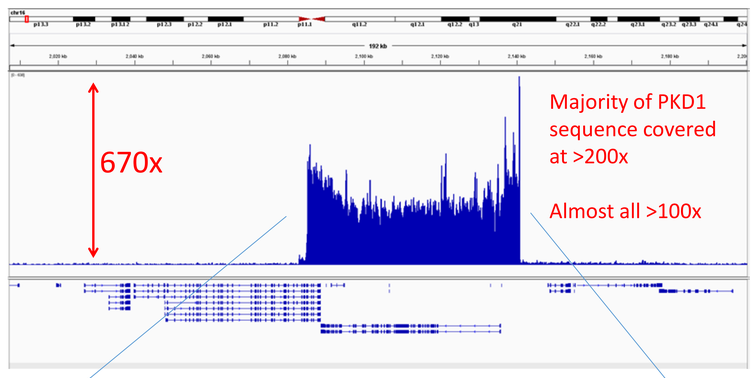
1. Identify target-enriched fractions by qPCR
2. Shear to 300bp with Covaris
3. Ethanol precipitate
4. Library construction with Kapa HyperPrep Library kit
5. QC with Bioanalyzer, Kapa Illumina Library Quant kit
6. Miseq, 2x150 bp PE sequencing
7. Data analysis (BWA MEM, Samtools, Bedtools, IGV)

Sequencing Results

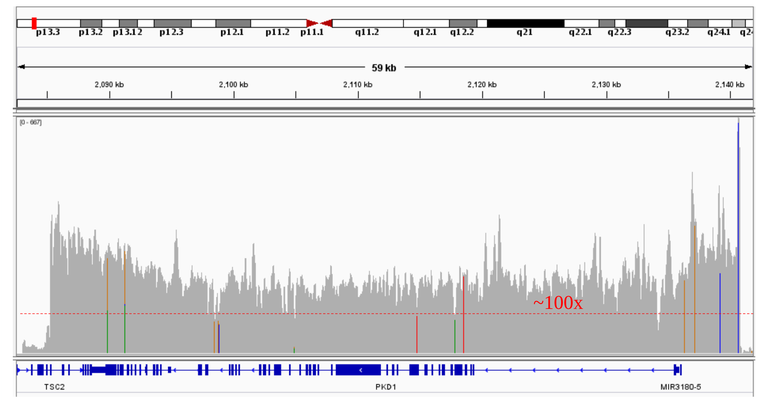
Bedgraph coverage over entire chr16



Bedgraph coverage near PKD1



IGV view of coverage and variants



Acknowledgements

We thank **Dr. Marjorie Beggs** of Arkana Laboratories (Little Rock, AR) for drawing our attention to the PKD1 gene as a medically important gene that could be addressed with the HLS-CATCH workflow.