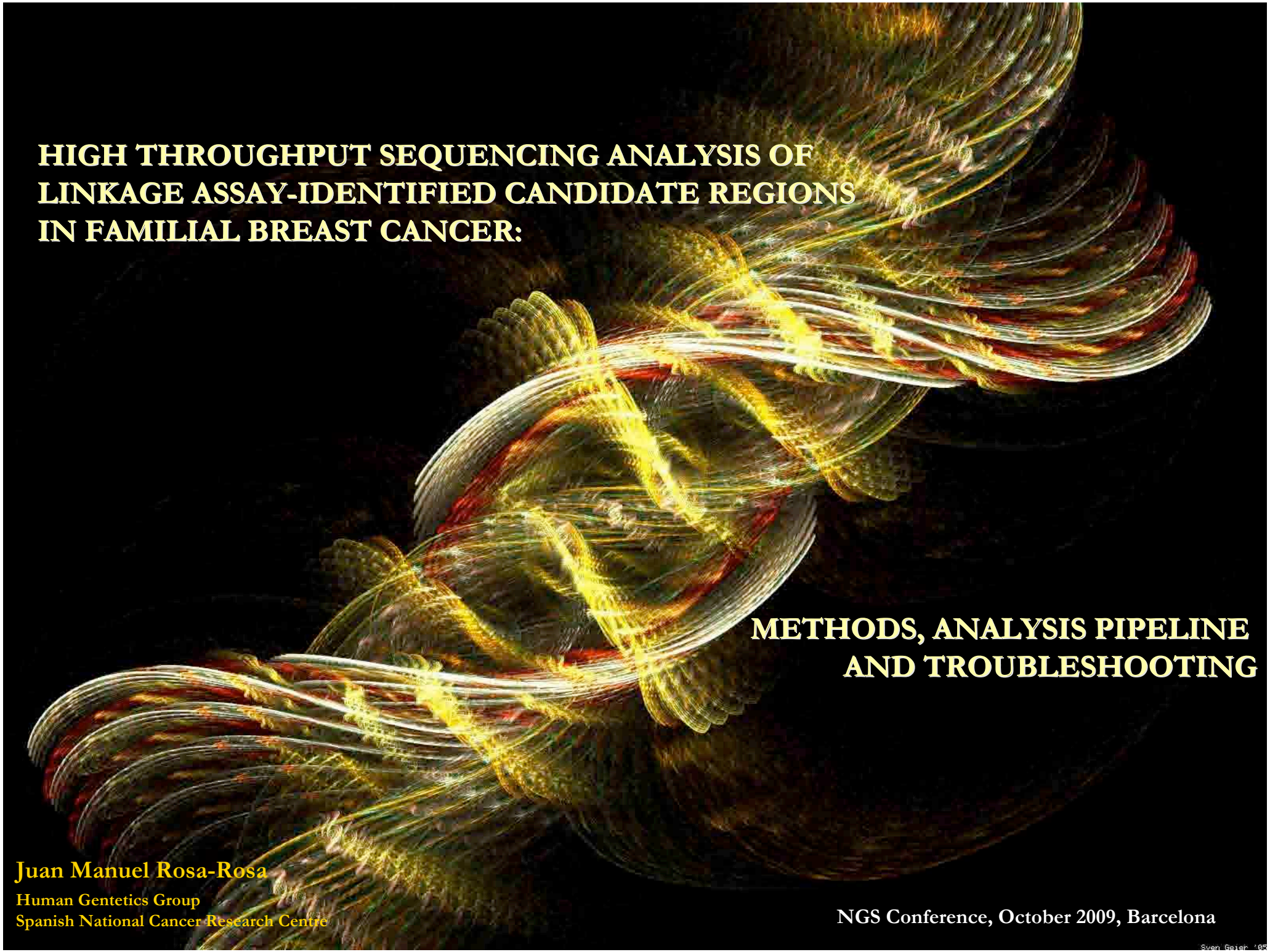


HIGH THROUGHPUT SEQUENCING ANALYSIS OF LINKAGE ASSAY-IDENTIFIED CANDIDATE REGIONS IN FAMILIAL BREAST CANCER:

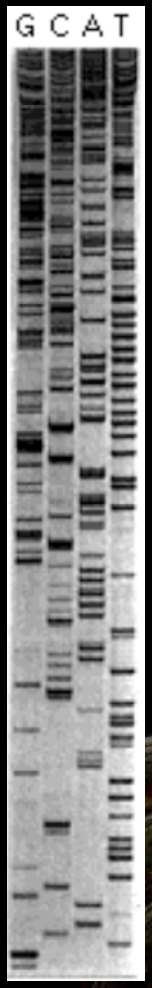
METHODS, ANALYSIS PIPELINE AND TROUBLESHOOTING

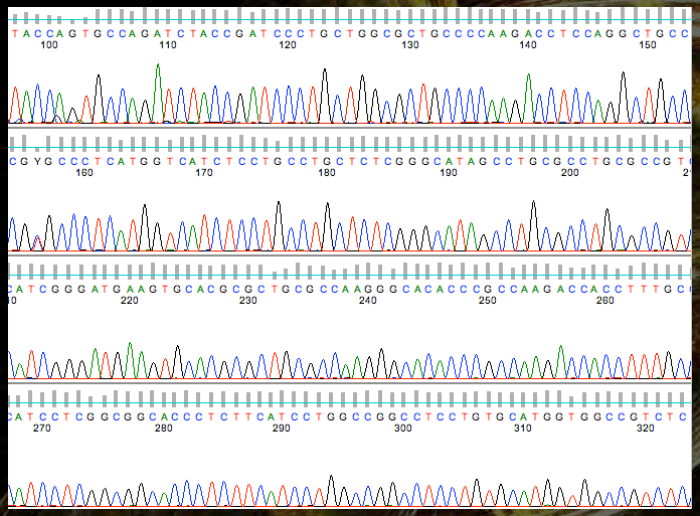
Juan Manuel Rosa-Rosa
Human Gentetics Group
Spanish National Cancer Research Centre

NGS Conference, October 2009, Barcelona

Deep sequencing is ….

… EVOLUTION …





```
@HANNIBAL_1_FC30ACCAAXX:2:1:14:776
AGCAGCATCATTCATAATACCCAAAACGTAG
+
>AB@>@BB@2?BBBBBB@9AB?2<<@<2,6?!
@HANNIBAL_1_FC30ACCAAXX:2:1:14:774
AGCAGCATCATTCATAATACCCAAAACGTAG
+
=BBBBBBB;@ABBB=BBBBBB???BBBBB33!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:1095
AGTGCTATGATTACAGGTGTGAGCCACTGCG
+
BBBBBCBBBBBCBB@>9ABBB>>>BBBB@9>!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:1081
AAAAGGACTTACCAATGATAGAAAAATTGCT
+
=BBB9<@B@?<???=B?2?@3'9B>;>>>@B!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:1387
AAAACCCACTTTCCCCATTTGCTCTGTAAAT
+
9<BBBBBBA9ABBBBBBBBBB?+<BBBB?=?!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:1712
AATGGAATGGAATGGAATGGAATGGAATGGA
+
BBA<*<CA>;9BA;><?B;9?<47@BBB6'6!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:747
ATATGATTCATCTGTTAGTTGTCACAAAATA
+
B6,9'.9B2;@BBB@=.,;6',6BB??<BBB!
@HANNIBAL_1_FC30ACCAAXX:2:1:15:618
CCTGTAATTCCAGCTACTCGTGAGGCTGAGG
```

… however, evolution has some costs …

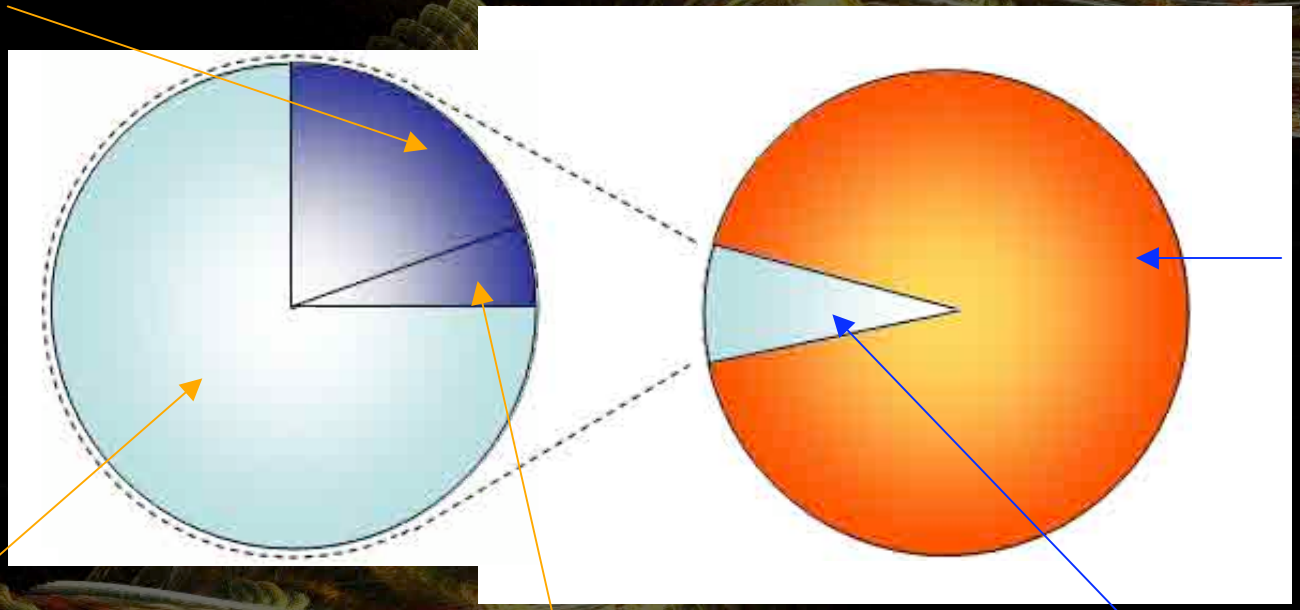
- Breast carcinoma is the first cause of cancer mortality among women worldwide (~ 1 million cases/year)

Familial Breast Cancer

All Breast Cancer

High susceptibility to breast cancer genes

BRCA1 / BRCA2



90-95%

Sporadic Breast Cancer

~70% →

Non-BRCA1/2 Families

Linkage Studies

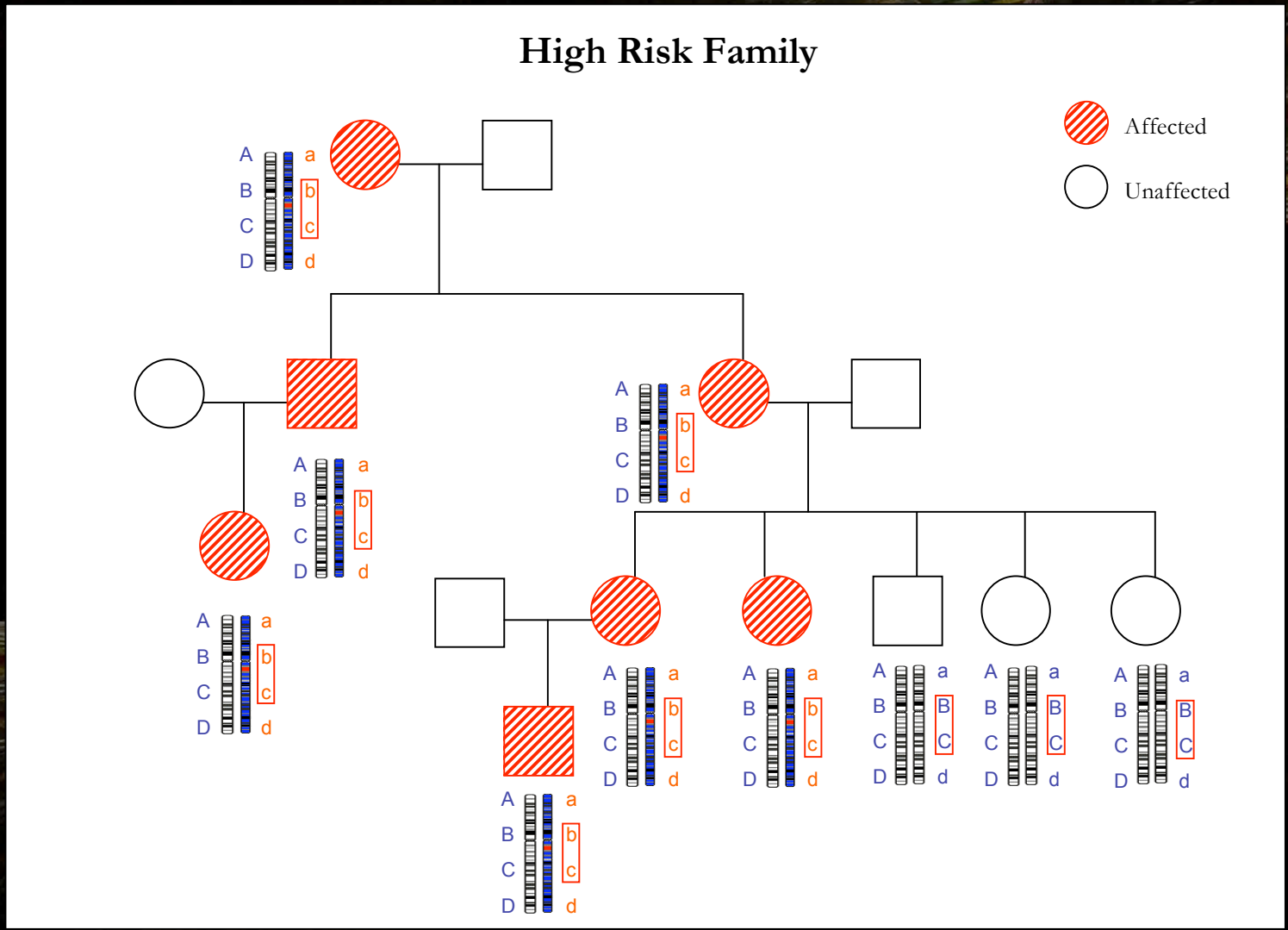
Other susceptibility genes

CHECK2, PALB2, p53, STK11, PTEN, ATM

5-10%

Familial or Hereditary Breast Cancer

Concept of Affected-Haplotype Sharing



Linkage Study on *non-BRCA1/2* families

Samples

132 samples from

41 *non-BRCA1/2* families

Methods

5800 SNPs (Linkage Panel 4.0)

1 SNP / 500 kb

Results

chr3 & chr6: *suggestive* linkage

chr21: *significant* linkage

Chromosome	Region	From	To	NPL(Max)	p value	Par Dom
3	q25.33-q26.2	rs1472578	rs1920122	2.46	0.007	**3.01**
6	q24.3-q25.1	rs612928	rs1407491	2.65	0.004	**2.26**
21	q22.13	rs1012959	rs2836301	4.37	0.00001	**3.55**

Families selected as putatively linked to each candidate region :

chr 3 → 6 families

chr 6 → 5 families

chr21 → 5 families

Rosa-Rosa et al. Am J Hum Genet, 2009

Exon-capture assay

Nature Genetics 39, 1522 - 1527 (2007)
Published online: 4 November 2007 | doi:10.1038/ng.2007.42

Genome-wide in situ exon capture for selective resequencing

Emily Hodges[1,4], Zhenyu Xuan[1,2,4], Vivekanand Balija[2], Melissa Kramer[2], Michael N Molla[3], Steven W Smith[3], Christina M Middle[3], Matthew J Rodesch[3], Thomas J Albert[3], Gregory J Hannon[1] & W Richard McCombie[2]

960 | VOL.4 NO.6 | 2009 | NATURE PROTOCOLS

Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing

Emily Hodges[1,2], Michelle Rooks[1,2], Zhenyu Xuan[1], Arindam Bhattacharjee[3], D Benjamin Gordon[3], Leonardo Brizuela[3], W Richard McCombie[1] & Gregory J Hannon[1,2]

Enrichment of specific sequences through CGH tiling arrays for selective resequencing

A pilot project linking CNIO and CSHL

- Mutational screening in the coding sequence of all the genes located on both *suggestive* candidate regions (chromosomes 3 and 6) identified in our linkage study

- Practical application of exon-capture assay

- Improvement in the data analysis pipeline

Samples

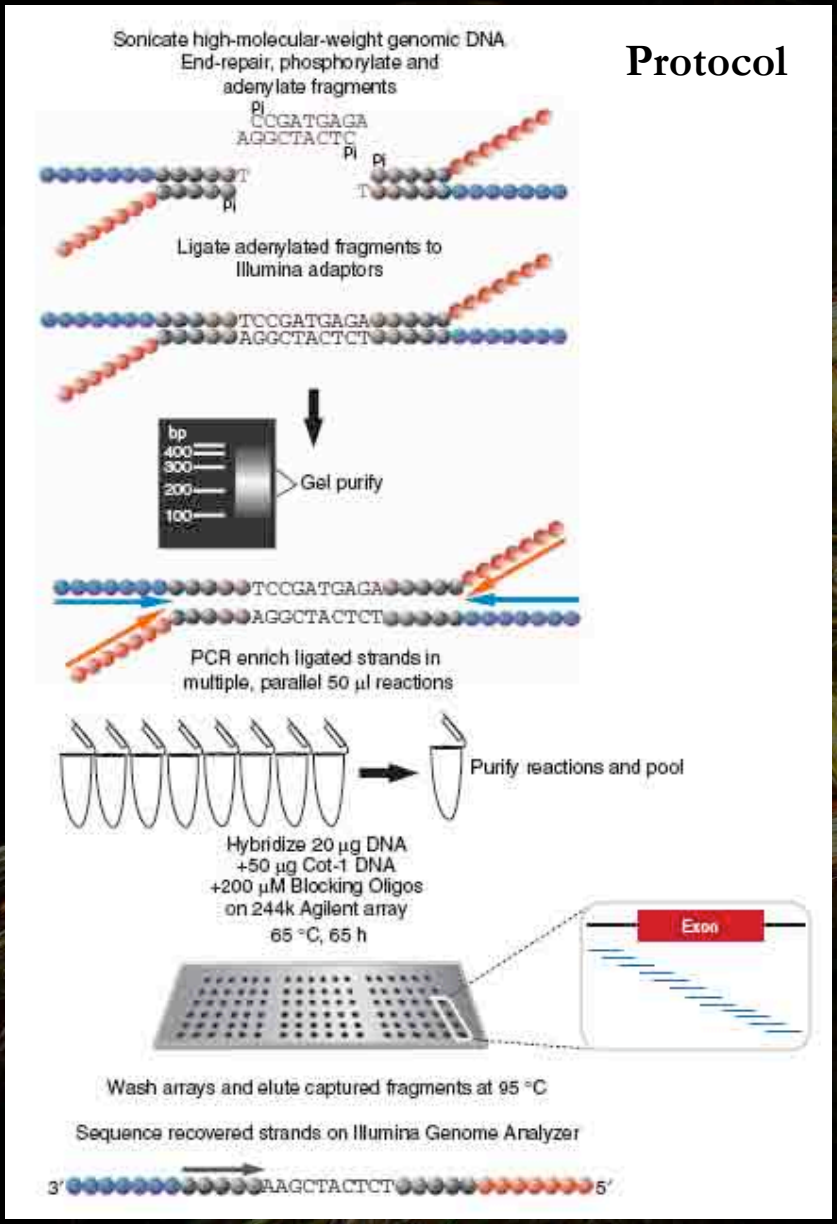
- DNA from 20 affected members from 9 different *non-BRCA1/2* families were collected
  At least two individuals per family to allow intrafamilial comparison

- DNA from 4 individuals from control population were pooled
  Many advantages related to sample homogeneity and heterozigosity

Regions

- Region on chr 3: 10 Mb
- Region on chr 6: 6 Mb
- Coding Regions from 159 genes ~ 400.000 bp

Protocol



a) **Library construction**
- Sonication
- End-repair
- Adaptor ligation
- PCR amplification

b) **Hybridization**

c) **Sequencing**
- Solexa 36 bp single-ends

Hodges et al. Nat Prot, 2009

Nature. 2008 November 6; 456(7218): 66–72. doi:10.1038/nature07485.

DNA sequencing of a cytogenetically normal acute myeloid leukemia genome

Timothy J Ley[1,2,3,4,*], Elaine R Mardis[2,3,*], Li Ding[2,3], Bob Fulton[3], Michael D McLellan[3], Ken Chen[3], David Dooling[3], Brian H Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E Larson[3], Dan C Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W Hillier[5], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R Osborne[3], Patrick Minx[3], David Gordon[5], Asif Chinwalla[3], Yu Zhao[1], Rhonda E Ries[1], Jacqueline E Payton[6], Peter Westervelt[1,4], Michael H Tomasson[1,4], Mark Watson[3,4,6], Jack Baty[7], Jennifer Ivanovich[4,8], Sharon Heath[1,4], William D Shannon[1,4], Rakesh Nagarajan[4,6], Matthew J Walter[1,4], Daniel C Link[1,4], Timothy A Graubert[1,4], John F DiPersio[1,4], and Richard K Wilson[2,3,4]

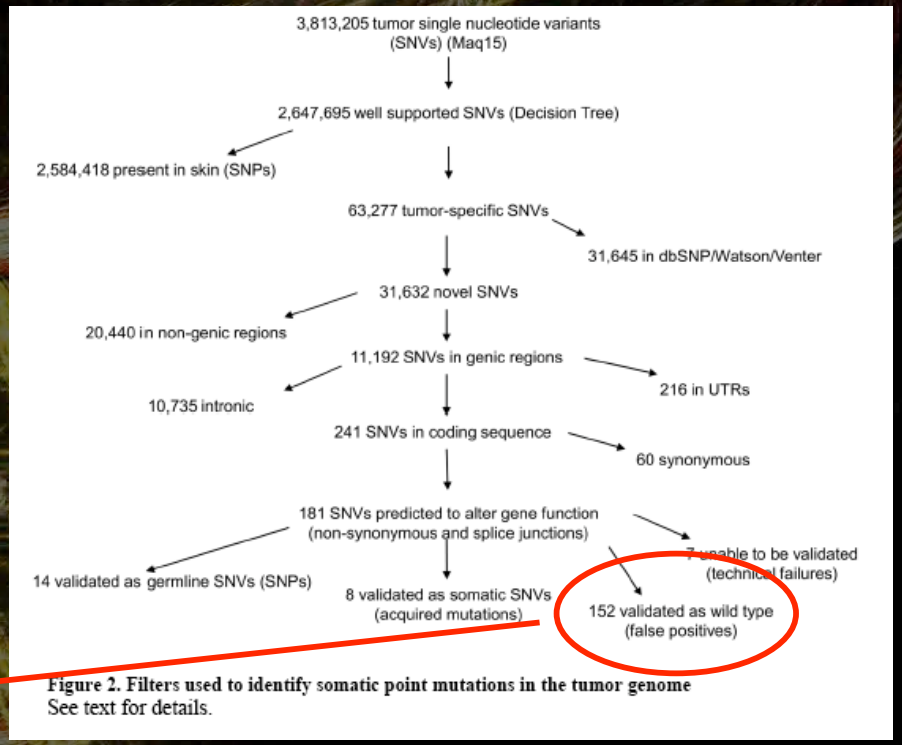


Figure 2. Filters used to identify somatic point mutations in the tumor genome
See text for details.

- False Positive Rate > 85%

… optimization was required

Alignment with SOAP v 1.0

- Indels identification in single-ends data
- Very informative output file

Scores

$$score = \left(\frac{X_V}{X_R}\right) \times 100$$

Quality Score

Xy = Average value of base-calling quality in variant alleles

Xr = Average value of base-calling quality in reference alleles

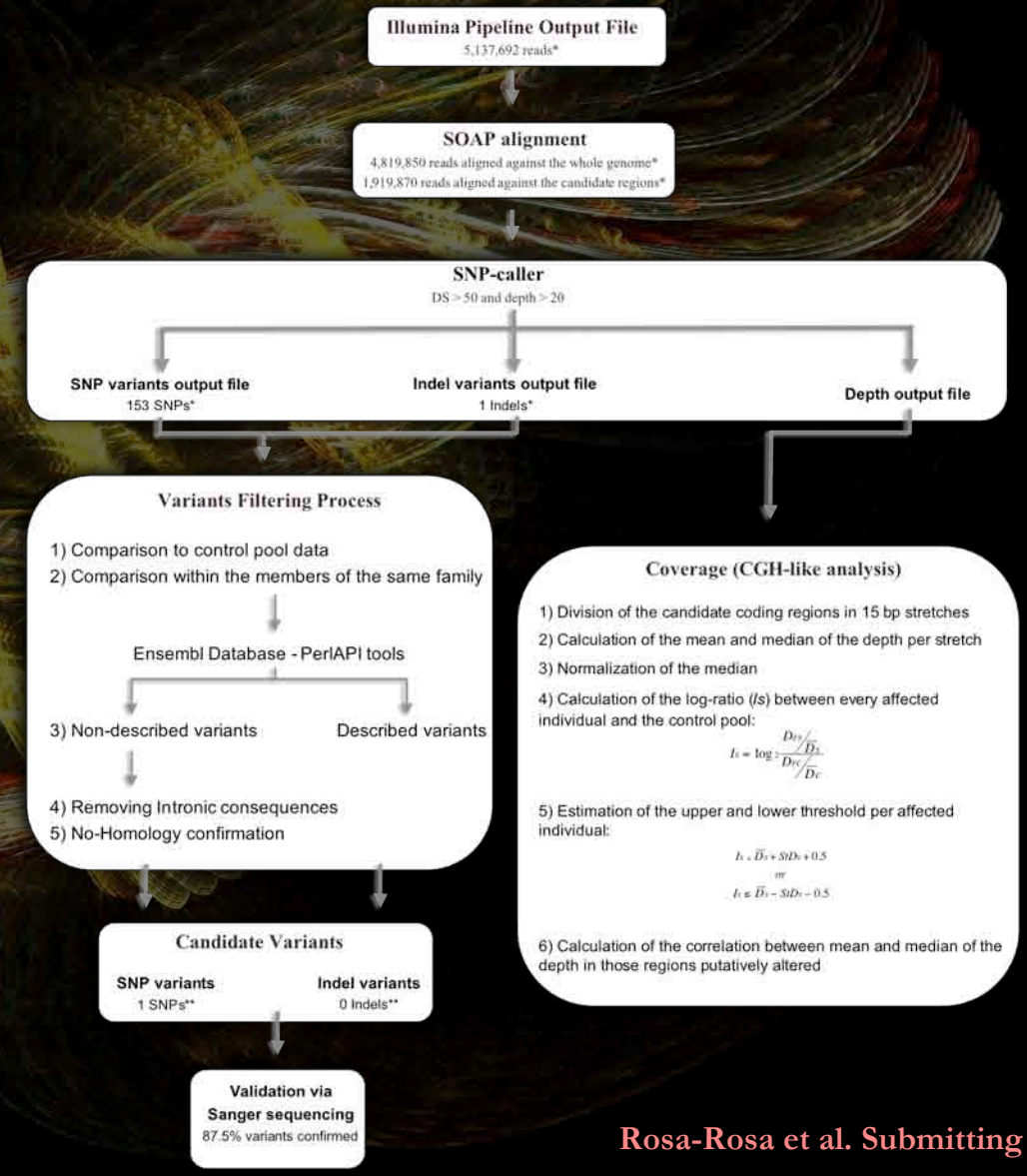
Depth Score

Xy = Depth value for variant allele

Xr = Depth value for reference allele

Scores always highlighted the relationship between variant and reference alleles

Analysis Pipeline



Rosa-Rosa et al. Submitting

Table 1

Chromosome	Family	Individual	Number of sequences: Total	Aligned to whole genome (%*)	Aligned to candidate regions (%**)	Depth: Coverage in %	Mean	Median
3	27	07S722	3,123,937	2,956,483 (94.64)	1,186,611 (40.14)	98.04	26	25
		07S723	4,922,157	4,538,392 (92.20)	1,518,625 (33.46)	98.43	29	29
		07S724	4,183,568	3,954,837 (94.53)	1,515,614 (38.32)	97.89	28	26
		07S725	2,952,969	2,839,271 (96.15)	1,168,679 (41.16)	97.11	24	22
	60	06-240	2,652,926	2,580,914 (97.29)	882,837 (34.21)	97.96	22	20
		96-652	5,934,453	4,737,175 (79.82)	1,670,157 (35.26)	98.15	28	24
	531	I-1408	12,228,047	11,188,204 (91.50)	4,694,871 (41.96)	99.07	57	48
		I-904	4,293,087	3,585,982 (83.53)	1,531,322 (42.70)	97.50	30	22
	713	07S635	7,568,672	7,442,938 (98.34)	2,793,056 (37.53)	99.11	45	44
		07S636	7,160,552	6,889,152 (96.21)	2,574,119 (37.36)	98.94	43	42
6	11	04-168	5,734,052	5,599,100 (97.65)	2,459,740 (43.93)	98.57	43	42
		96-265	6,240,024	6,012,522 (96.35)	2,642,942 (43.96)	98.22	35	32
	40	07S576	2,006,661	1,667,648 (83.11)	779,723 (46.76)	97.11	18	17
		07S581	4,016,214	3,618,178 (90.09)	1,568,060 (43.34)	97.66	25	23
	929	I-1627	5,811,276	5,665,182 (97.49)	2,311,149 (40.80)	98.52	33	32
		I-3345	2,602,250	2,554,051 (98.15)	1,059,131 (41.47)	98.27	23	23
	990	I-1927	8,134,956	7,903,785 (97.16)	3,029,994 (38.34)	98.84	51	50
		I-1928	7,922,500	7,590,406 (95.81)	2,817,358 (37.12)	99.02	49	48
	1125	I-2033	2,747,911	2,666,280 (97.03)	1,105,059 (41.45)	97.87	24	23
		I-4347	2,517,619	2,406,505 (95.59)	1,088,350 (45.23)	97.74	24	24
		TOTAL	102.753.831	96,397,005 (93.81)	38,397,397 (39.83)			
		Average Aff	5,137,692	4,819,850 (93.63)	1,919,870 (40.22)	98.20	33	31
		Control pool	22,390,251	18,221,565 (81.38)	7,438,610 (40.82)	99.33	111	98
		Average All	5,214,336	4,775,773 (91.58)	1,909,833 (39.99)	98.25	37	34

Rosa-Rosa et al. Submitting

Table 2

Chr	Family	Individual	SNPs	After control	Shared by family	Non-described	Consequences	Exonic	Candidate SNPs (%)*
3	27	07S722	102	2 2	0	0	0	0	0 (0.00)
		07S723	93	9					
		07S724	120	3 2					
		07S725	84	1 9					
	60	06-240	78	19	1 0	5	4	3	1 (10.00)
		96-652	96	2 6					
	531	I-1408	95	2 0	5	2	1	1	0 (0.00)
		I-904	109	3 8					
	713	07S635	110	3 8	1 2	5	6	3	1 (8.33)
		07S636	111	3 0					
6	11	96_265	179	5 3	1 5	3	8	2	0 (0.00)
		04_168	182	5 2					
	40	07S581	226	8 9	4 3	1 1	29	22	2 (4.65)
		07S576	257	1 4 0					
	929	I_3345	175	4 5	1 8	5	13	10	0 (0.00)
		I_1627	198	6 6					
	990	I_1927	246	9 6	5 1	1 7	56	23	5 (9.80)
		I_1928	231	8 2					
	1125	I_4347	185	5 3	1 9	3	10	0	0 (0.00)
		I_2033	181	6 6					
		Average	153	50	19	6	14	7	1 (3.62)

Rosa-Rosa et al. Submitting

Table 3

Chr	Family	Position (hg18)	Gen	Reference	Variant	QS[a]	DS[b]	Consequence[c]		Alamuth prediction[d]	Gene function
3	60	161301596	AC026118.17	A	T	91/91	56/57	NCG		-	pseudogene
	713	170284589	EVI1	A	G	95/94	128/100	3UTR		-	hematopoietic proliferation protein, related to acute myeloid leukemia
6	40	151203125	PLEKHG1	C	T	103/98	133/146	SYN	S1186S	-	unknown
		151713613	AKAP12	C	T	98/96	185/183	SYN	P700P	-	scaffold protein in signal transduction, is a cell growth-related protein
	990	146761618	GRM1	C	T	101/96	101/81	NSYN	R584C	AFF	metabotropic glutamate receptor
		150087915	NUP43	T	C	95/93	101/97	3UTR		-	part of a nuclear pore complex, mediating bidirectional transport of macromolecules between the cytoplasm and nucleus
		150205485	LRP11	T	C	101/101	74/95	NSYN	I312V	NDB	unknown
		151564223	AL451072.14	G	A	93/98	119/116	NCG		-	non-coding RNA

- Confirmation rate = 0.875

Rosa-Rosa et al. Submitting

Table 4

Chr	Fam	Ind	Mean	St Dev	Upper	Lower
3	27	07S722	-0.13	0.75	1.12	-1.38
		07S723	-0.04	0.31	0.77	-0.86
		07S724	-0.07	0.35	0.77	-0.92
		07S725	-0.10	0.43	0.82	-1.03
	60	06-240	-0.01	0.40	0.89	-0.91
		69-652	0.00	0.59	1.10	-1.09
	531	I-1408	0.05	0.53	1.08	-0.97
		I-904	-0.31	1.83	2.02	-2.64
	713	07S635	-0.04	0.62	1.08	-1.16
		07S636	-0.04	0.45	0.92	-0.99
6	11	04-168	-0.09	0.35	0.76	-0.94
		96-265	-0.02	0.36	0.83	-0.88
	40	07S576	-0.05	0.73	1.18	-1.29
		07S581	0.00	0.41	0.91	-0.91
	929	I-1627	-0.02	0.36	0.83	-0.88
		I-3345	-0.09	0.36	0.77	-0.95
	990	I-1927	-0.08	0.80	1.22	-1.38
		I-1928	-0.03	0.63	1.10	-1.16
	1125	I-2033	-0.03	0.44	0.91	-0.97
		I-4347	-0.09	0.39	0.80	-0.98
		Global	-0.06	0.55	0.99	-1.11

No bias in global coverage

Rosa-Rosa et al. Submitting

In summary …

• … we designed an analysis pipeline for mutational screening via SOAP v1.0 that resulted in a low false positive rate with a low probability of discarding real positive variants

• … we identified seven variants that passed all the different filters to be considered candidates, however further functional studies are required to assess whether any of them is an actual causal mutation or a polymorphism

• … we regard the present strategy as a valid second step after linkage studies in order to identify candidate high penetrance genes

Acknowledgements

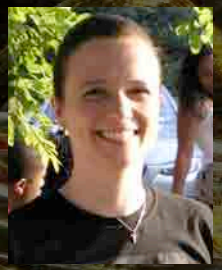
Human Genetics Group - CNIO

Javier Benitez

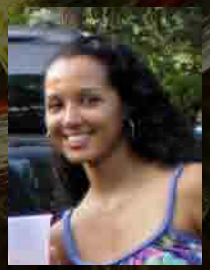
Javier Gracia
Guillermo Pita

CSHL

Greg Hannon



Emily Hodges



Michelle Rooks

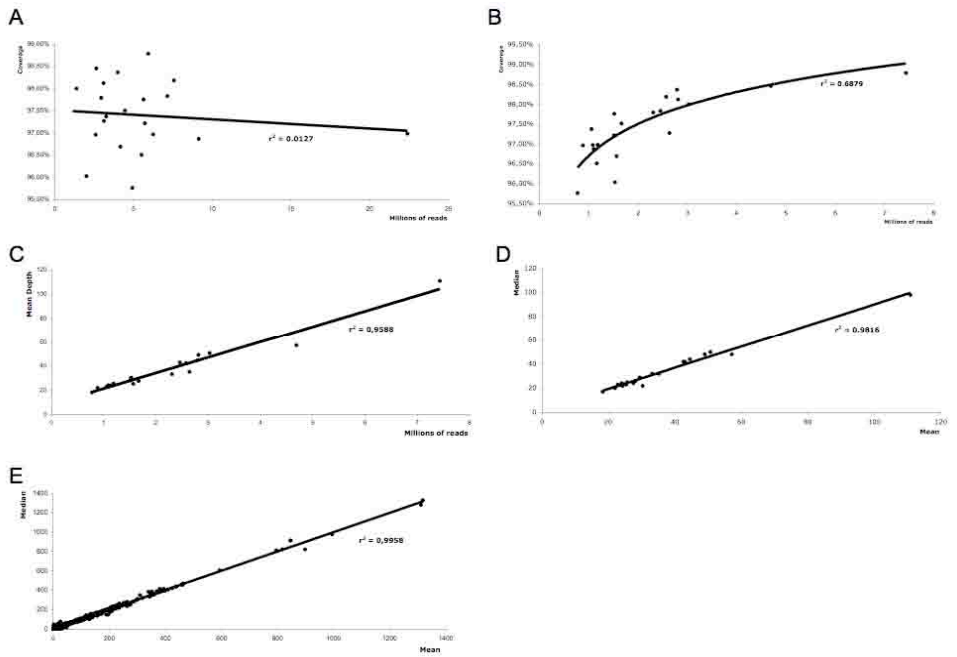
Irving Cancer Research

Jose Silva

Supplementary Information

Family	Individual	Chromosome	Position	Gen	Reference allele	Variant allele	Genotype	QS	DS	Consequence	Global Depth
60	06_240	3	162442989	NMD3	G	T	G/T	80	50	ESSENTIAL_SPLICE_SITE	9
	06_240	3	166262966	SI	C	A	C/A	73	50	ESSENTIAL_SPLICE_SITE	3
	06_240	3	168534468	ZBBX	C	A	C/A	81	50	ESSENTIAL_SPLICE_SITE	3
	06_240	6	146281062	SHPRH	C	A	C/A	88	100	ESSENTIAL_SPLICE_SITE	2
	06_240	6	147039141	C6orf103	G	T	G/T	77	57	ESSENTIAL_SPLICE_SITE	11
	06_240	6	147625118	STXBP5	G	T	G/T	79	66	ESSENTIAL_SPLICE_SITE	10
	06_240	6	147872067	SAMD5	C	T	C/T	77	50	STOP_GAINED	3
	06_240	6	151228808	MTHFD1L	C	T	C/T	105	50	STOP_GAINED	3
531	I_904	3	171238698	GPR160	G	T	G/T	110	50	ESSENTIAL_SPLICE_SITE	3
	I_904	6	148706054	SASH1	T	C	T/C	59	100	ESSENTIAL_SPLICE_SITE	2
	I_904	6	151857015	C6orf97	T	G	T/G	76	50	ESSENTIAL_SPLICE_SITE	3
	I_904	3	161426731	AC112641.3	C	A	C/A	117	100	STOP_GAINED	2
27	07S722	3	161601302	SMC4	A	C	A/C	107	50	ESSENTIAL_SPLICE_SITE	3
	07S722	3	161601303	SMC4	G	C	G/C	92	50	ESSENTIAL_SPLICE_SITE	3
	07S722	3	162305379	B3GALNT1	T	G	T/G	54	50	ESSENTIAL_SPLICE_SITE	6
	07S723	3	162303593	B3GALNT1	A	G	A/G	56	100	ESSENTIAL_SPLICE_SITE	2
	07S723	3	166263999	SI	G	A	G/A	102	50	STOP_GAINED	3
	07S725	3	161483097	IFT80	T	A	T/A	65	50	ESSENTIAL_SPLICE_SITE	3
	07S725	3	166197219	SI	C	A	C/A	72	50	ESSENTIAL_SPLICE_SITE	3
	07S725	3	171460450	PRKCI	G	T	G/T	89	50	ESSENTIAL_SPLICE_SITE	6
	07S725	6	146177430	FBXO30	A	T	A/T	59	50	ESSENTIAL_SPLICE_SITE	3
	07S725	6	146308467	SHPRH	C	A	C/A	75	53	ESSENTIAL_SPLICE_SITE	23
	07S725	6	150001390	KATNA1	C	A	C/A	80	57	ESSENTIAL_SPLICE_SITE	11
	07S725	3	168566393	ZBBX	G	T	G/T	81	133	STOP_GAINED	7
	07S725	3	169247461	GOLIM4	C	A	C/A	78	50	STOP_GAINED	3
	07S725	6	147007611	C6orf103	G	T	G/T	74	50	STOP_GAINED	3
11	96_265	3	161477961	IFT80	C	A	C/A	57	50	ESSENTIAL_SPLICE_SITE	3
40	07S576	3	166247495	SI	C	A	C/A	71	100	ESSENTIAL_SPLICE_SITE	8
	07S576	3	168534468	ZBBX	C	A	C/A	129	50	ESSENTIAL_SPLICE_SITE	3
	07S576	6	149680846	MAP3K7IP2	T	G	T/G	66	50	ESSENTIAL_SPLICE_SITE	3
	07S576	6	151228880	MTHFD1L	T	G	T/G	107	100	ESSENTIAL_SPLICE_SITE	2
	07S576	3	168728494	WDR49	G	T	G/T	56	50	STOP_GAINED	3
	07S581	6	150251497	RAET1E	C	A	C/A	108	100	ESSENTIAL_SPLICE_SITE	2
	07S581	3	168560381	ZBBX	C	A	C/A	130	100	STOP_GAINED	2
	07S581	6	150506161	PPP1R14C	C	A	C/A	68	50	STOP_GAINED	3
990	I_1927	6	149680846	MAP3K7IP2	T	G	T/G	85	50	ESSENTIAL_SPLICE_SITE	3
	I_1927	6	151754247	ZBTB2	C	A	C/A	88	50	ESSENTIAL_SPLICE_SITE	3
1125	I_2033	6	147625118	STXBP5	G	T	G/T	68	66	ESSENTIAL_SPLICE_SITE	5
	I_2033	6	147672912	STXBP5	G	T	G/T	68	100	ESSENTIAL_SPLICE_SITE	4
	I_4347	3	161736380	KPNA4	C	A	C/A	69	50	ESSENTIAL_SPLICE_SITE	4
	I_4347	3	166247495	SI	C	A	C/A	73	50	ESSENTIAL_SPLICE_SITE	6
	I_4347	6	147677092	STXBP5	G	T	G/T	59	100	ESSENTIAL_SPLICE_SITE	2

Family	Individual	Chromosome	Position (hg18)	Gen	Reference allele	Variant allele	Genotype	QS	DS	Consequence	Global Depth
21	05_980	3	160965047	SCHIP1	C	A	C/A	98	26	STOP_GAINED	34
		3	166192869	SI	C	A	C/A	71	26	STOP_GAINED	19
		6	146797238	GRM1	C	A	C/A	65	33	STOP_GAINED	4
		6	151754247	ZBTB2	C	A	C/A	69	25	ESSENTIAL_SPLICE_SITE	6
		6	151754247	ZBTB2	C	T	C/T	74	25	ESSENTIAL_SPLICE_SITE	6



Supplementary Figure 1: Correlations.

The coverage along the candidate regions was very high (98% on average) and no correlation between it and the number of sequences obtained per individual was observed (A), although we observed a logarithmic trend when the number of sequences aligned to the candidate regions was used (B). On the other hand, a strong correlation between the number of sequences aligned to the candidate coding regions and the mean depth was observed in our dataset (C). Failures in the capture step were discarded since high correlations between the global mean and the global median of the depth per individual (D) and between the mean and the median of the depth in putatively altered 15-bp regions for all the individuals (E) were observed (see text for details).