Detection of *hydin* Gene Duplication in Personal Genome Sequence Data

Jong-Il Kim[1,2], Young Seok Ju[1,3], Sheehyun Kim[4], Dongwan Hong[1] and Jeong-Sun Seo[1,2,3,4]*

[1]Genomic Medicine Institute, Medical Research Center, Seoul National University, Seoul 110-799, Korea, [2]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, Korea, [3]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, Korea, [4]Macrogen Inc., Seoul 153-023, Korea

Abstract

Human personal genome sequencing can be done with high efficiency by aligning a huge number of short reads derived from various next generation sequencing (NGS) technologies to the reference genome sequence. One of the major obstacles is the incompleteness of human reference genome. We tried to analyze the effect of hidden gene duplication on the NGS data using the known example of *hydin* gene. *Hydin2*, a duplicated copy of *hydin* on chromosome 16q22, has been recently found to be localized to chromosome 1q21, and is not included in the current version of standard human genome reference. We found that all of eight personal genome data published so far do not contain *hydin2*, and there is large number of nsSNPs in *hydin*. The heterozygosity of those nsSNPs was significantly higher than expected. The sequence coverage depth in *hydin* gene was about two fold of average depth. We believe that these unique finding of *hydin* can be used as useful indicators to discover new hidden multiplication in human genome.

Keywords: gene duplication, *hydin*, next generation sequencing technology, personal genome

Next generation sequencing (NGS) technology has made personal genome sequencing possible in the laboratory scale by decreasing the cost and time for genome sequencing (Tucker *et al*., 2009; Yngvadottir, 2009). We have recently finished the whole genome sequencing of a Korean male (AK1) (Kim *et al*., 2009) and a Korean female (AK2) (unpublished). At least five more personal genomes have also been reported to be sequenced by NGS technology (Ahn *et al*., 2009; Bentley *et al*., 2008; McKernan *et al*., 2009; Wang *et al*., 2008; Wheeler *et al*., 2008) One of the major drawbacks in NGS technology is relative short read length compared to the conventional capillary sequencing. For this reason, the read sequences have to be aligned to one reference sequence rather than being assembled with each others. The most frequently used reference is the human reference genome sequence derived from the Human Genome Project (IHGSC, 2004). Although this reference sequence have been constantly updated and used by number of researchers world-wide , these sequences still have large number of gaps, and are far from being completed. Therefore, if a part of the sample genome is largely different from the reference or if counterpart is missing in the reference genome, analysis of this part by NGS may be impossible or likely to be erroneous. In this paper, we tried to study the effect of hidden duplication of human *hydin* in the reference genome on NGS data.

Mouse *hydin* was first identified as a candidate gene in the Hy3 mouse model of hydrocephalus (Davy and Robinson, 2003). Doggett *et al*. found that human *hydin* gene, the ortholog of murine *hydin*, had been duplicated from chromosome 16q22.2 into chromsosome 1q21.1 (Doggett *et al*., 2006). In addition to its localization, the sequences, exon/intron structure and expression pattern based on the EST profile of *hydin2* have been identified and are available in public database (http://www.ncbi.nlm.nih.gov). However, the exact location of *hydin2* insertion in chromosome 1 is still uncertain. Therefore, build 37, the current version of human reference genome does not contain *hydin2* in the sequence of chromosome 1.

During the analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in AK1 and AK2, we found that *hydin* is one of the so-called super-SNP genes, which have much higher number of nsSNPs than other genes. When we compared the number of nsSNPs in *hydin* of other personal genome sequences, all had similarly large number of nsSNPS (Table 1). The average nsSNP density of *hydin* in eight genomes (1.66 ± 0.34 per kb coding sequence) were significantly higher than the average genome-wide nsSNP density (0.27 ± 0.02, $p < 10^{-5}$). This finding is less likely to be originated from technical problem in sequencing or analysis procedures, considering that all different genome data showed sim-

*Corresponding author: E-mail jeongsun@snu.ac.kr
Tel +82-2-740-8246, Fax +82-2-741-5423
Accepted 12 September 2009

Table 1. Number of nsSNPs in eight different personal genome sequence data

Genome D	Platform	# of nsSNP in *hydin*	nsSNP density* (number/kb)	Homozygous nsSNP	Heterozygous nsSNP	References
AK1	GA	35	1.88	1	34	Kim et al., 2009
SJK	GA	39	2.09	1	38	Ahn et al., 2009
AK2	SOLiD	28	1.50	3	25	Unpublished
YH	GA	32	1.72	1	31	Wang et al., 2008
Yoruba1[†]	GA	38	2.04	1	37	Bentley et al., 2008
Yoruba2	SOLiD	24	1.29	1	23	McKeman et al., 2009
Watson	GS-20	30	1.61	0	30	Wheeler et al., 2008
Venter	AB3730	21	1.13	1	20	Levy et al., 2007
Total[‡]		54	2.90			

*Number of nsSNPs per 1 kb cds (coding sequence) of *hydin*.

[†] This paper used two different methods for aligning. We used the result made by MAQ.

[‡] Total number of nsSNP sites was calculated by counting the nsSNPs occurring at the same position in different data as one site.

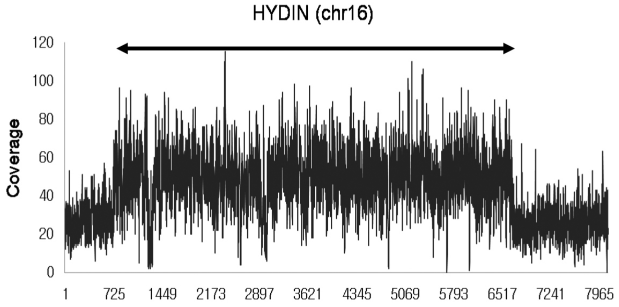

Fig. 1. Coverage depth profile of AK1 genome in part of chrosome 16 which contains *hydin* (marked by an arrow). Horizontal axis is relative position (base) in this region, and vertical axis is the coverage depth of sequence reads aligned to each position.

ilar increase in the number of nsSNPs.

Another clue explaining this finding was the ratio of heterozygous and homozygous nsSNPs in *hydin* (Table 1). Heterozygous and homozygous SNPs were 62.5% and 37.5% of total SNPs in AK1. This pattern was similar in nsSNPs, 61.7% and 38.3% of total nsSNPs being heterozygous and homozygous nsSNPs, respectively. In *hydin* of AK1, however, only one of 35 nsSNP was homozygous. Other seven genome data showed almost same ratio of homozygous and heterozygous nsSNPs.

All these finding might be explained by the gene duplication. Because there is no *hydin2* in the reference genome, all the sequence reads derived from *hydin2* as well as *hydin* had to be aligned to only *hydin* of the reference genome. This will make the alignment result as if *hydin* of these samples were tetraploidy, that is four copies, being double of other areas in autosome. To confirm this hypothesis, we examined coverage depth profile of *hydin* in AK1 genome. As clearly seen in Fig. 1, the average coverage depth of *hydin* gene is roughly 60x, being double of 30x in other regions.

Finally we compared each variant sequences determined as nsSNPs in *hydin* with the known sequence of *hydin2*. The variant alleles of all 35 nsSNPs found in *hydin* of AK1 with the exception of only one, were found to be same to the wild type alleles at the counterpart position of *hydin2* (data not shown). Because we can not determine whether it comes from *hydin* or *hydin2* when a sequence read is aligned to *hydin* in reference genome, these data do not guarantee that the difference between *hydin* and *hydin2* could be the only origin of the enrichment of nsSNP genes. Further sequencing analysis using clone of cDNA and/or genomic DNA from AK1 will be required for complete understanding of more detailed structure of both *hydin* and *hydin2*. We believe that at least some part of the other super-SNP genes found in AK1 or other genomes are originated by the hidden duplication like *hydin*. Our result suggest that the criteria we found in *hydin*: 1) enrichment of nsSNPs, 2) increased ratio of heterozygous to homozygous SNPs, and 3) increase in coverage depth, can be used as useful indicators to discover the hidden structural variation in human genomes.

Acknowledgements

This work was supported by a grant (KRF-2005-041-E00049) from the Korean Ministry of Education, Science and Technology.

References

Ahn, S.M., Kim, T.H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C., Park, D., Lee, Y.S., Kim, S., Reja, R., Jho, S., Kim, C.G., Cha, J.Y., Kim, K.H., Lee, B., Bhak, J., Kim, S.J. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* 19, 1622-1629.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes Fajardo, K.V., Scott Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A., Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., and Smith, A.J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53-59.

Davy, B.E., and Robinson, M.L. (2003). Congenital hydrocephalus in hy3 mice is caused by a frameshift mutation in *hydin*, a large novel gene. *Hum. Mol. Genet.* 12, 1163-1170.

Doggett, N.A., Xie, G., Meincke, L.J., Sutherland, R.D., Mundt, M.O., Berbari, N.S., Davy, B.E., Robinson, M.L., Rudd, M.K., Weber, J.L., Stallings, R.L., and Han, C. (2006). A 360-kb interchromosomal duplication of the human *hydin* locus. *Genomics* 88, 762-771.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945.

Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J., Kim, H.S., Chung, I.S., Lee, W.C., Lee, J.S., Seo, S.H., Yun, J.Y., Woo, H.N., Lee, H., Suh, D., Lee, S., Kim, H.J., Yavartanoo, M., Kwak, M., Zheng, Y., Lee, M.K., Park, H., Kim, J.Y., Gokcumen, O., Mills, R.E., Zaranek, A.W., Thakuria, J., Wu, X., Kim, R.W., Huntley, J.J., Luo, S., Schroth, G.P., Wu, T.D., Kim, H., Yang, K.S., Park, W.Y., Kim, H., Church, G.M., Lee, C., Kingsmore, S.F., and Seo J.S. (2009). A highly annotated whole-genome sequence of a Korean individual. *Nature* 460, 1011-1015.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L., and Venter, J.C. (2007). The diploid genome sequence of an individual human. *PLoS. Biol.* 5, e254.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M., and Blanchard, A.P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19, 1527-1541.

Tucker, T., Marra, M., and Friedman, J.M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *Am. J. Hum. Genet.* 85, 142-154.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., and Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature* 456, 60-65.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A., and Rothberg, J.M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876.

Yngvadottir, B., Macarthur, D.G., Jin, H., and Tyler-Smith, C. (2009). The promise and reality of personal genomics. *Genome Biol.* 10, 237.