Assessing Population Based Differences of Average Total Depth of Coverage in Next Generation Sequencing

Table 1: Population Specific Mean Depth of Coverage across all chromosomes for Yoruban, African (YRI) and Central European (CEU) populations in the 1000 Genomes Phase 1 Pilot Study Exon Capture Sequencing Project (targeted depth of 50x) and Low-Coverage Whole Genome Sequencing (target depth of 2-4x) in the 1000 Genomes Phase 1 pilot study.

	Exon Capture Sequencing Project				**Low-Coverage Whole Genome Sequencing Project**			
	CEU		**YRI**		**CEU**		**YRI**	
	N=36		N=25		N=91		N=119	
Chr	Number of Variants	Total Depth of Coverage Mean (STD)	Number of Variants	Total Depth of Coverage Mean (STD)	Number of Variants	Total Depth of Coverage Mean (STD)	Number of Variants	Total Depth of Coverage Mean (STD)
1	344	83.6 (33.1)	505	62.0 (31.2)	3,325,487	7.8 (1.5)	6,935,828	8.2 (1.7)
2	237	84.8 (33.2)	330	66.4 (27.1)	1,049,521	7.9(1.4)	3,633,300	8.2(1.7)
3	176	88.6 (35.4)	262	68.8 (28.0)	556,362	8.0 (1.4)	3,008,969	8.2 (1.7)
4	146	77.2 (30.2)	241	62.6 (24.5)	567,547	7.8(1.3)	773,805	7.8(1.7)
5	168	79.2 (33.4)	250	54.2 (27.2)	499,164	8.0(1.4)	696,878	8.1(1.7)
6	253	66.8 (37.9)	341	53.9 (31.7)	518,645	7.94(1.4)	687,537	8.1(1.7)
7	145	72.8(36.2)	430	53.0 (25.0)	451,004	7.9(1.4)	610,887	8.1(1.7)
8	125	77.6(30.1)	201	62.5 (24.4)	429,055	8.0(1.4)	605,510	8.1(1.7)
9	116	73.5 (40.3)	183	52.8 (29.8)	328,069	8.0(1.4)	450,846	8.2(1.7)
10	182	95.9 (47.9)	250	72.9 (33.0)	396,487	8.0(1.4)	530,275	8.3(1.7)
11	196	96.5 (40.8)	337	75.2 (38.4)	381,826	8.0(1.4)	518,208	8.3(1.7)
12	185	74.0 (30.9)	302	56.5 (26.7)	365,883	8.0(1.4)	493,878	8.2(1.7)
13	59	87.6 (34.5)	64	74.4 (32.3)	293,253	7.8(1.4)	389,516	7.9(1.7)
14	139	82.7 (41.7)	218	61.6 (31.4)	254,837	7.8(1.4)	344,054	8.2(1.7)
15	168	83.6 (34.8)	228	63.1 (27.7)	210,540	7.6(1.3)	298,300	8.3(1.7)
16	139	67.0 (33.2)	185	46.0 (26.6)	238,117	7.7(1.5)	334,638	8.6(1.8)
17	182	72.6 (33.1)	255	57.1 (28.0)	196,327	7.6(1.5)	815,928	8.6(1.7)
18	113	74.9 (39.8)	161	62.0 (32.7)	225,279	7.6(1.3)	309,033	8.1(1.7)
19	210	59.7 (37.1)	310	43.5 (28.7)	157,182	7.3(1.6)	209,370	8.5(1.8)
20	116	56.6 (39.8)	154	47.4 (33.7)	174,484	7.8(1.4)	244,110	8.6(1.7)
21	23	69.1 (27.0)	42	50.6 (16.0)	109,143	7.5(1.5)	152,505	8.1(1.8)
22	67	60.2 (34.1)	111	41.7 (26.5)	101,568	7.6(1.6)	137,859	8.7(1.7)

CEU= The Central European population identified in the 1000 genomes dataset; YRI= The Yoruban Nigerian population identified in the 1000 Genomes dataset. Chr= chromosome; N= Total number of individuals in the populations used for analysis; # of Variants = total number of variants in the population that were detected at each chromosome; average coverage is based on the mean total depth of coverage from all individuals within the population

Table 2. The association between average total depth of coverage and population group (CEU and YRI) from the exon capture and low-coverage whole genome phase 1 pilot study datasets in the 1000 Genomes Project.

	Exon Capture Sequencing Project				**Low-Coverage Whole Genome Sequencing Project**			
	CEU	**YRI**	**Statistical Comparisons**		**CEU**	**YRI**	**Statistical Comparisons**	
	N=91	N=119	T-Test	Kolmogorov-Smirnov	N=36	N=25	T-Test	Kolmogorov-Smirnov
Chr	# of Variants	# of Variants	Mean Difference	Ksa	# of Variants	# of Variants	Mean Difference	Ksa
1	344	505	21.7[a]	1.9	3,325,487	6,935,828	-0.4[a]	905.9
2	237	330	18.4[a]	0.7	1,049,521	3,633,300	-0.4[a]	559.4
3	176	262	19.8[a]	0.6	556,362	3,008,969	-0.2[a]	456.8
4	146	241	14.6[a]	0.7	567,547	773,805	-0.02[a]	387.3
5	168	250	25.0[a]	1.7	499,164	696,878	-0.1[a]	360.3
6	253	341	12.9[a]	0.8	518,645	687,537	-0.2[a]	353.0
7	145	430	19.9[a]	1.7	451,004	610,887	-0.2[a]	328.4
8	125	201	15.1[a]	0.5	429,055	605,510	-0.2[a]	334.0
9	116	183	20.7[a]	1.0	328,069	450,846	-0.2[a]	283.6
10	182	250	22.4[a]	0.6	396,487	530,275	-0.3[a]	302.5
11	196	337	21.3[a]	1.4	381,826	518,208	-0.3[a]	299.5
12	185	302	17.5[a]	0.9	365,883	493,878	-0.2[a]	297.2
13	59	64	13.3[b]	0.1	293,253	389,516	-0.1[a]	270.7
14	139	218	21.1[a]	1.1	254,837	344,054	-0.4[a]	233.0
15	168	228	20.5[a]	1.1	210,540	298,300	-0.7[a]	204.1
16	139	185	20.9[a]	1.4	238,117	334,638	-1.0[a]	197.1
17	182	255	15.6[a]	0.6	196,327	815,928	-0.9[a]	200.7
18	113	161	12.9[b]	0.4	225,279	309,033	-0.5[a]	218.1
19	210	310	16.2[a]	1.3	157,182	209,370	-1.1[a]	125.0
20	116	154	9.20[b]	0.7	174,484	244,110	-0.8[a]	178.8
21	23	42	18.5[a]	1.2	109,143	152,505	-0.5[a]	141.7
22	67	111	18.5[a]	1.5	101,568	137,859	-1.1	67.50

CEU= The Central European population identified in the 1000 genomes dataset; YRI= The Yoruban Nigerian population identified in the 1000 Genomes dataset. Chr= chromosome; N= sample size; Ksa= the Kolmogorov-Statistic; # of Variants = total number of variants in the population that were detected at each chromosome; average coverage is based on the mean total depth of coverage from all individuals within the

population; [a] Results are statistically significant using a p-value of 0.002 which corrects for multiple comparisons (0.05/22).; [b] Results are marginally significant using a p-value of 0.05.

References

1. Adeyemo, A., & Rotimi, C. (2014). What does genomic medicine mean for diverse populations?. *Molecular genetics & genomic medicine*, *2*(1), 3-6.
2. Ajay, S.S., Parker, S.C., Abaan, H.O., Fajado, K.V., Marguilies, E.H. Accurate and comprehensive sequencing of personal genomes. *Genomes Research.* 21, 1498-1505.
3. Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., & Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human genetics*, *131*(10), 1541-1554.
4. Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 11-10.
5. Bentley, D. et al (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456; 53-59.
6. Biesecker, L. G., Burke, W., Kohane, I., Plon, S. E., & Zimmern, R. (2012). Next-generation sequencing in the clinic: are we ready?. *Nature Reviews Genetics*, *13*(11), 818-824.
7. Cohen, N., Dagan, T., Stone, L., & Graur, D. (2005). GC composition of the human genome: in search of isochores. *Molecular biology and evolution*, *22*(5), 1260-1272.
8. Curtis, K., Talwalkar, A., Zaharia, M., Fox, A., & Patterson, D. A. (2015). SiRen: Leveraging Similar Regions for Efficient & Accurate Variant Calling.
9. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, *43*(5), 491-498.
10. DFCI. Dana Farber Cancer Institute. (2015). Joint Center for Cancer Precision Medicine.: Dana-Farber, Birgham and Women's Boston's Children's Hospital and Broad Institute to collaborate to harness genetic data for patients. November 12, 2013. Retrieved From: http://www.dana-farber.org/Newsroom/News-Releases/joint-center-for-cancer-precision-medicine-established.aspx
11. Gafni, E., Luquette, L. J., Lancaster, A. K., Hawkins, J. B., Jung, J. Y., Souilmi, Y., ... & Tonellato, P. J. (2014). COSMOS: Python library for massively parallel workflows. *Bioinformatics*, btu385.
12. Genome Analysis Toolkit (GATK) (2014). Best Practice Variant Detection with the GATK v4, for release 2.0. Retrieved from http://gatkforums.broadinstitute.org/discussion/1186/best-practice-variant-detection-with-the-gatk-v4-for-release-2-0-retired
13. Genome Analysis Toolkit (GATK) (2014). HC overview: How the HaplotypeCaller works. Retrieved from https://www.broadinstitute.org/gatk/guide/article?id=4148
14. Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301-304.
15. IHGSC-International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931-945.
16. Institute of Medicine. (2012). Genome-Based Therapeutics: Targeted Drug Discovery and Development: Workshop Summary. National Academies Press(US).
17. Jarow, J. (2013). Consortium sequences, assembles Human Genome Project donor to improve reference assembly. Genome web. Retrieved from https://www.genomeweb.com/sequencing/consortium-sequences-assembles-human-genome-project-donor-improve-reference-asse

18. Jorde, L. B., & Wooding, S. P. (2004). Genetic variation, classification and'race'. *Nature genetics*, *36*, S28-S33.
19. Krimsky, S. (2012). The short life of a race drug. *The Lancet*, *379*(9811), 114-115.
20. Lander, E.S., Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* 2; 231-239.
21. Lee, S. S. J., & Mudaliar, A. (2009). Racing forward: the genomics and personalized medicine act. *Science (New York, NY)*, *323*(5912), 342.
22. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589-595.
23. Li, R., Li, Y., Zheng, H., Luo, R., Zhu, H., Li, Q., ... & Wang, J. (2010). Building the sequence map of the human pan-genome. *Nature biotechnology*, *28*(1), 57-6
24. McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., ... & Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics*, *40*(10), 1166-1174.
25. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.
26. Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, *11*(1), 31-46.
27. Motulsky, A. G. (2010). History of Human Genetics*. In *Vogel and Motulsky's Human Genetics* (pp. 13-29). Springer Berlin Heidelberg.
28. National Institutes of Health. Biological Sciences Curriculum Study. NIH Curriculum Supplement Series. Retrieved from: http://www.ncbi.nlm.nih.gov/books/NBK20363/
29. Need, A. C., & Goldstein, D. B. (2009). Next generation disparities in human genomics: concerns and remedies. *Trends in Genetics*, *25*(11), 489-494.
30. Nielsen, R. (2004). Population genetic analysis of ascertained SNP data.*Human genomics*, *1*(3), 218-224.
31. Ramos, E., Callier, S. L., & Rotimi, C. N. (2012). Why personalized medicine will fail if we stay the course. *Personalized medicine*, *9*(8), 839-847.
32. Rehm, H., Bale, S., Bayrak-Toydemir, B., Berg, J., Brown, K., Deignan, J., Fries, M., Funke, B., Hegde, M., Lyon, E., ACMG Laboratory Quality Assurance Committee. ACMG Clinical Laboratory Standards for Next-Generation Sequencing.
33. Robasky, K, Lewis, N.E., Church, G. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56-62.
34. Rosenfeld, J. A., Mason, C. E., & Smith, T. M. (2012). Limitations of the human reference genome for personalized genomics. *PLoS One*, *7*(7), e40294.
35. Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature*, *200*(8), 16-18.
36. Shaw, J. (2015). Toward Precision Medicine. *Harvard Magazine,* May-June. Retreived From: http://harvardmagazine.com/2015/05/toward-precision-medicine
37. Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses.*Nature Reviews Genetics*, *15*(2), 121-132.

38. Thomas, U.G. (2014). Community-wide Effort Aims to Better Represent Variation in Human Reference Genome. Genome web. Retrieved from https://www.genomeweb.com/informatics/community-wide-effort-aims-better-represent-variation-human-reference-genome.
39. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., ... & Abecasis, G. R. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics*, *46*(4), 409-415.
40. Warden, C. D., Adamson, A. W., Neuhausen, S. L., & Wu, X. (2014). Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, *2*, e600.
41. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., ... & Rothberg, J. M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *nature*, *452*(7189), 872-876.
42. Witherspoon, D. J., Wooding, S., Rogers, A. R., Marchani, E. E., Watkins, W. S., Batzer, M. A., & Jorde, L. B. (2007). Genetic similarities within and between human populations. *Genetics*, *176*(1), 351-359.
43. 1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061-1073.