the WINNOWER

Proof of Extensive Copy Number Variation in The Human Genome

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. 2006. "Global variation in copy number in the human genome." Nature no. 444 (7118):444-54. doi: 10.1038/nature05329.

STEPHEN W. SCHERER [1]

1. Hospital for Sick Children and University of Toronto, Canada

CORRESPONDENCE:
stephen.scherer@sickkids.ca

DATE RECEIVED:
September 03, 2014

DOI:
10.15200/winn.140076.67673

KEYWORDS:
cnv, copy number, genome, variation

A primary message from the Human Genome Project, as well as from earlier studies, was that DNA in the genomes of any two individuals is 99.9 per cent identical. The 0.1 per cent variation was attributed to some three million single nucleotide polymorphisms (SNPs) scattered amongst the chromosomes. Larger genomic changes -- involving losses or gains of thousands or millions of nucleotides, encompassing one or several genes -- were thought to be exceptionally rare, and almost always involved in disease (Check 2005).

In the summer of 2004, Charles Lee, a Canadian cytogeneticist at Harvard, collaborating with my Toronto team, and (separately) Michael Wigler's group at Cold Spring Harbor, published data showing that numerous larger gene-sized (or greater) segments of DNA were present in different copy numbers in all individuals (Iafrate et al. 2004, Sebat et al. 2004). These studies, which used microarrays to scan the genomes of a handful of individuals, found about a dozen novel copy number variations (CNVs) per genome. Within the clinical genetics community, uptake of the concept of CNVs was immediate, and as part of our original paper (Iafrate et al. 2004) we established a database (now called the *Database of Genomic Variants*) to make our results accessible and applicable. The genome sequencing community, however, was busy finishing a "one-size-fits-all' sequence map (Human Genome Sequencing 2004)and starting their venture to map common SNPs, (the HapMap Project) (2005), and did not yet fully recognize the potential importance of the CNV discoveries.

Given the limited resolution of the technologies used for the 2004 papers, and the small number of samples analyzed, we recognized that our initial CNV findings were likely just "the tip of the iceberg'. With simple extrapolations from our existing data, we predicted that every genome would likely have thousands of CNVs, as well as insertions and deletions (called indels, which are smaller CNVs (1 to 1,000 nucleotides)). We began to design experiments that might allow us to discover these. As we drafted the first grants in early to mid-2004, we quickly anticipated the challenges ahead: (i) selecting the most appropriate technology (s) for robust CNV detection, (ii) identifying appropriate control

samples from world populations, and (iii) securing funding for what would surely be a multi-million dollar CNV genome project. Charles and I found a champion for the project in microarray expert, Nigel Carter (and newly hired scientist Matthew Hurles) at the Welcome Trust Sanger Institute in the United Kingdom, and I was able to win big grants from the Canadian Institutes of Health Research and Genome Canada. We decided to study DNA samples from the same 270 individuals as used in the international HapMap project; these originated from four populations with ancestry in Europe, Africa or Asia.

To detect CNVs, we used two complementary genome-wide technologies. The first compared each sample to a reference standard, looking for differences in intensity among a set of more than 26,000 large-insert cloned segments of DNA called bacterial artificial chromosomes. This was done on microarrays that spanned nearly all of the euchromatic genome. The second was a proprietary genotyping approach developed by Keith Jones' team at Affymetrix. Some 500,000 SNPs were assayed, looking for stretches of adjacent SNP loci with atypical or unexpected allele ratios. For this purpose, Hiro Aburatani's group at the University of Tokyo helped by developing new computer algorithms for "calling" CNVs from SNP genotypes.

Coverage with these combined microarray approaches allowed detection of CNVs larger than 10kb, and we identified a total of 1,447 alterations among the 270 HapMap samples. CNV regions were estimated to involve an average, per genome, of more than 20 million base pairs -- some 5- to 10-fold more variation per nucleotide than had been suggested by studying SNPs alone. In fact, it was startling to discover that, in the 270 human samples tested, 12% of the genome was copy number variable. These regions encompass about 2,900 genes, or 10% of those known. Some CNVs in the general population are millions of bases in size, involving many genes, yet sometimes with no observable consequence. We also found CNVs that involved genes known to be disease-associated, indicating these genomic events could have significant clinical consequences. We karyotyped all of the cell lines, examined the genetic relationship of SNPs and CNVs, and, with Chris Tyler-Smith of the Sanger Institute and Don Conrad (then a student with Jonathan Pritchard at the University of Chicago), studied properties of CNVs in population and evolutionary genetics. Given all of the striking data generated by our study, we argued strongly, both in the abstract and discussion of the publication (Redon et al. 2006), that CNV assessment should be standard in the design of all studies of the genetic basis of phenotypic variation, including those about disease susceptibility or evolution. In *Nature's* accompanying *News & Views* section, geneticist Huntington Willard wrote, "the data suggest that the greatest source of genetic diversity in our species lies not in the millions of SNPs, but rather in larger segments of the genome whose presence or absence calls into question what exactly is a "normal' human genome" (Shianna and Willard 2006).

The major obstacle we encountered in publishing this work was the enormity of methods and data generated. The human/clinical genetics community anxiously awaited public release of the data, to enable their own research into genetic diseases. With great effort, we released all of our data at many public sites, including the *Database of Genomic Variants*, and assembled a massive 233 page Supplemental section in the *Nature* paper itself. We also coordinated the same-day publication of two accompanying technical papers in *Genome Research,* detailing the clone-based microarrays and computational tools (Fiegler et al. 2006) and the Affymetrix SNP microarrays and calling algorithms (Komura et al. 2006). Lars Feuk, in my group, took the lead to publish a fourth paper on the same day, in *Nature Genetics* (Khaja et al. 2006), describing our use of genome assembly comparison to identify smaller CNVs and indels, and balanced alterations like inversions. This set the stage for subsequent copy number and structural variation maps of the human genome, with even higher resolution.

In recollection, I believe that we assembled just the right team of junior and senior scientists, who not only recognized the importance of CNV very early on, but also had "fire-in-the-belly' to meet what often were seen as unobtainable goals and timelines. Our core group (Sanger Institute-Lee-Scherer labs) continued to work together and used experience from this foundational study to produce a second-generation CNV map of the genome with higher resolution (Conrad et al. 2010), and to enable studies

of the impact of CNVs on gene expression (Stranger et al. 2007), genetic disease (Craddock et al. 2010), and in personal genome analysis (Park et al. 2010).

I suppose that when Charles and I originally contemplated what eventually would become widely dubbed as the "Redon *et al Nature* CNV study", it was simply to add proof of the existence of CNVs, as recognized in 2004, and to show that it was a ubiquitous form of natural genetic variation in humans. This massive study of human genetic variation tied together (for the first time) cytogenetics, submicroscopic copy number variation, and single nucleotide polymorphisms, providing a framework and standard for future studies of the human genome. Ultimately, the sheer magnitude of CNV discovery -- the essence of which we captured in the title "*Global variation in copy number in the human genome*'-- is what has made this paper a citation classic.

REFERENCES

2005. "A haplotype map of the human genome." *Nature* no. 437 (7063):1299-320. doi: 10.1038/nature04226.

Check, E. 2005. "Human genome: patchwork people." *Nature* no. 437 (7062):1084-6. doi: 10.1038/4371084a.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. 2010. "Origins and functional impact of copy number variation in the human genome." *Nature* no. 464 (7289):704-12. doi: 10.1038/nature08516.

Craddock, N., M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, C. Holmes, J. L. Marchini, K. Stirrups, M. D. Tobin, L. V. Wain, C. Yau, J. Aerts, T. Ahmad, T. D. Andrews, H. Arbury, A. Attwood, A. Auton, S. G. Ball, A. J. Balmforth, J. C. Barrett, I. Barroso, A. Barton, A. J. Bennett, S. Bhaskar, K. Blaszczyk, J. Bowes, O. J. Brand, P. S. Braund, F. Bredin, G. Breen, M. J. Brown, I. N. Bruce, J. Bull, O. S. Burren, J. Burton, J. Byrnes, S. Caesar, C. M. Clee, A. J. Coffey, J. M. Connell, J. D. Cooper, A. F. Dominiczak, K. Downes, H. E. Drummond, D. Dudakia, A. Dunham, B. Ebbs, D. Eccles, S. Edkins, C. Edwards, A. Elliot, P. Emery, D. M. Evans, G. Evans, S. Eyre, A. Farmer, I. N. Ferrier, L. Feuk, T. Fitzgerald, E. Flynn, A. Forbes, L. Forty, J. A. Franklyn, R. M. Freathy, P. Gibbs, P. Gilbert, O. Gokumen, K. Gordon-Smith, E. Gray, E. Green, C. J. Groves, D. Grozeva, R. Gwilliam, A. Hall, N. Hammond, M. Hardy, P. Harrison, N. Hassanali, H. Hebaishi, S. Hines, A. Hinks, G. A. Hitman, L. Hocking, E. Howard, P. Howard, J. M. Howson, D. Hughes, S. Hunt, J. D. Isaacs, M. Jain, D. P. Jewell, T. Johnson, J. D. Jolley, I. R. Jones, L. A. Jones, G. Kirov, C. F. Langford, H. Lango-Allen, G. M. Lathrop, J. Lee, K. L. Lee, C. Lees, K. Lewis, C. M. Lindgren, M. Maisuria-Armer, J. Maller, J. Mansfield, P. Martin, D. C. Massey, W. L. McArdle, P. McGuffin, K. E. McLay, A. Mentzer, M. L. Mimmack, A. E. Morgan, A. P. Morris, C. Mowat, S. Myers, W. Newman, E. R. Nimmo, M. C. O'Donovan, A. Onipinla, I. Onyiah, N. R. Ovington, M. J. Owen, K. Palin, K. Parnell, D. Pernet, J. R. Perry, A. Phillips, D. Pinto, N. J. Prescott, I. Prokopenko, M. A. Quail, S. Rafelt, N. W. Rayner, R. Redon, D. M. Reid, Renwick, S. M. Ring, N. Robertson, E. Russell, D. St Clair, J. G. Sambrook, J. D. Sanderson, H. Schuilenburg, C. E. Scott, R. Scott, S. Seal, S. Shaw-Hawkins, B. M. Shields, M. J. Simmonds, D. J. Smyth, E. Somaskantharajah, K. Spanova, S. Steer, J. Stephens, H. E. Stevens, M. A. Stone, Z. Su, D. P. Symmons, J. R. Thompson, W. Thomson, M. E. Travers, C. Turnbull, A. Valsesia, M. Walker, N. M. Walker, C. Wallace, M. Warren-Perry, N. A. Watkins, J. Webster, M. N. Weedon, A. G. Wilson, M. Woodburn, B. P. Wordsworth, A. H. Young, E. Zeggini, N. P. Carter, T. M. Frayling, C. Lee, G. McVean, P. B. Munroe, A. Palotie, S. J. Sawcer, S. W. Scherer, D. P. Strachan, C. Tyler-Smith, M. A. Brown, P. R. Burton, M. J. Caulfield, A. Compston, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, C. G. Mathew, M. Pembrey, J. Satsangi, M. R. Stratton, J. Worthington, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. Ouwehand, M. Parkes, N. Rahman, J. A. Todd, N. J. Samani, and P. Donnelly. 2010. "Genome-wide

association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." *Nature* no. 464 (7289):713-20. doi: 10.1038/nature08979.

Fiegler, H., R. Redon, D. Andrews, C. Scott, R. Andrews, C. Carder, R. Clark, O. Dovey, P. Ellis, L. Feuk, L. French, P. Hunt, D. Kalaitzopoulos, J. Larkin, L. Montgomery, G. H. Perry, B. W. Plumb, K. Porter, R. E. Rigby, D. Rigler, A. Valsesia, C. Langford, S. J. Humphray, S. W. Scherer, C. Lee, M. E. Hurles, and N. P. Carter. 2006. "Accurate and reliable high-throughput detection of copy number variation in the human genome." *Genome Res* no. 16 (12):1566-74. doi: 10.1101/gr.5630906.

Human Genome Sequencing, ConsortiumInternational. 2004. "Finishing the euchromatic sequence of the human genome." *Nature* no. 431 (7011):931-945. doi: http://www.nature.com/nature/journal/v431/n7011/suppinfo/nature03001_S1.html.

Iafrate, A. J., L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee. 2004. "Detection of large-scale variation in the human genome." *Nat Genet* no. 36 (9):949-51. doi: 10.1038/ng1416.

Khaja, R., J. Zhang, J. R. MacDonald, Y. He, A. M. Joseph-George, J. Wei, M. A. Rafiq, C. Qian, M. Shago, L. Pantano, H. Aburatani, K. Jones, R. Redon, M. Hurles, L. Armengol, X. Estivill, R. J. Mural, C. Lee, S. W. Scherer, and L. Feuk. 2006. "Genome assembly comparison identifies structural variants in the human genome." *Nat Genet* no. 38 (12):1413-8. doi: 10.1038/ng1921.

Komura, Daisuke, Fan Shen, Shumpei Ishikawa, Karen R. Fitch, Wenwei Chen, Jane Zhang, Guoying Liu, Sigeo Ihara, Hiroshi Nakamura, Matthew E. Hurles, Charles Lee, Stephen W. Scherer, Keith W. Jones, Michael H. Shapero, Jing Huang, and Hiroyuki Aburatani. 2006. "Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays." *Genome Research* no. 16 (12):1575-1584. doi: 10.1101/gr.5629106.

Park, Hansoo, Jong-Il Kim, Young Seok Ju, Omer Gokcumen, Ryan E. Mills, Sheehyun Kim, Seungbok Lee, Dongwhan Suh, Dongwan Hong, Hyunseok Peter Kang, Yun Joo Yoo, Jong-Yeon Shin, Hyun-Jin Kim, Maryam Yavartanoo, Young Wha Chang, Jung-Sook Ha, Wilson Chong, Ga-Ram Hwang, Katayoon Darvishi, HyeRan Kim, Song Ju Yang, Kap-Seok Yang, Hyungtae Kim, Matthew E. Hurles, Stephen W. Scherer, Nigel P. Carter, Chris Tyler-Smith, Charles Lee, and Jeong-Sun Seo. 2010. "Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing." *Nat Genet* no. 42 (5):400-405. doi: http://www.nature.com/ng/journal/v42/n5/suppinfo/ng.555_S1.html.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. 2006. "Global variation in copy number in the human genome." *Nature* no. 444 (7118):444-54. doi: 10.1038/nature05329.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. 2004. "Large-scale copy number polymorphism in the human genome." *Science* no. 305 (5683):525-8. doi: 10.1126/science.1098918.

Shianna, K. V., and H. F. Willard. 2006. "Human genomics: in search of normality." *Nature* no. 444 (7118):428-9. doi: 10.1038/444428a.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. de Grassi, C. Lee, C. Tyler-Smith, N. Carter, S. W. Scherer, S. Tavare, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. 2007. "Relative impact of nucleotide and copy number variation on gene expression

phenotypes." *Science* no. 315 (5813):848-53. doi: 10.1126/science.1136678.