# REFERENCES (I)


Bekker, A., Holland, H.D., Wang, P.L., Rumble, D., 3rd, Stein, H.J., Hannah, J.L., Coetzee, L.L. and Beukes, N.J., 2004, "Dating the Rise of Atmospheric Oxygen", **Nature**, Vol. 427, No. 6970, pp. 117-120.

Bhaya, D., 2004, "Light Matters: Phototaxis and Signal Transduction in Unicellular Cyanobacteria", **Molecular Microbiology**, Vol. 53, No. 3, pp. 745-754.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A. and Andersson, S.G., 2004, "Computational Inference of Scenarios for Alpha-Proteobacterial Genome Evolution", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 101, No. 26, pp. 9722-9727.

Bryant, D.A. and Frigaard, N.U., 2006, "Prokaryotic Photosynthesis and Phototrophy Illuminated", **Trends in Microbiology**, Vol. 14, No. 11, pp. 488-496.

Buick, R., 2008, "When Did Oxygenic Photosynthesis Evolve?", **Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences**, Vol. 363, No. 1504, pp. 2731-2743.

Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. and Roos, D.S., 2006, "Orthomcl-Db: Querying a Comprehensive Multi-Species Collection of Ortholog Groups", **Nucleic Acids Research**, Vol. 34, No. Database issue, pp. D363-D368.

Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S., 2007, "Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes", **PLoS One**, Vol. 2, No. 4, p. e383.

Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M. and DeSalle, R., 2006, "Orthologid: Automation of Genome-Scale Ortholog Identification within a Parsimony Framework", **Bioinformatics**, Vol. 22, No. 6, pp. 699-707.

Cohen, Y., Jorgensen, B.B., Revsbech, N.P. and Poplawski, R., 1986, "Adaptation to Hydrogen Sulfide of Oxygenic and Anoxygenic Photosynthesis among Cyanobacteria", **Applied and Environmental Microbiology**, Vol. 51, No. 2, pp. 398-407.

Deluca, T.F., Wu, I.H., Pu, J., Monaghan, T., Peshkin, L., Singh, S. and Wall, D.P., 2006, "Roundup: A Multi-Genome Repository of Orthologs and Evolutionary Distances", **Bioinformatics**, Vol. 22, No. 16, pp. 2044-2046.

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K.V., Allen, J.F., Martin, W. and Dagan, T., 2008, "Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor", **Molecular Biology and Evolution**, Vol. 25, No. 4, pp. 748-761.

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T., de Marsac, N.T., Wincker, P., Dossat, C., Ferriera, S., Johnson, J., Post, A.F., Hess, W.R. and Partensky, F., 2008, "Unraveling the Genomic Mosaic of a Ubiquitous Genus of Marine Cyanobacteria", **Genome Biology**, Vol. 9, No. 5, p. R90.

Eisen, J.A., 1998, "Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis", **Genome Research**, Vol. 8, No. 3, pp. 163-167.

Eisen, J.A. and Fraser, C.M., 2003, "Phylogenomics: Intersection of Evolution and Genomics", **Science**, Vol. 300, No. 5626, pp. 1706-1707.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and et al., 1995, "Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd", **Science**, Vol. 269, No. 5223, pp. 496-512.

Garrity, G.M., Brenner, D.J., Krieg, N.R. and Staley, J.R., 2005, **Bergey's Manual of Systematic Bacteriology, Volume Two: The Proteobacteria, Parts a - C**, Springer - Verlag.

Guan, X., Qin, S., Zhao, F., Zhang, X. and Tang, X., 2007, "Phycobilisomes Linker Family in Cyanobacterial Genomes: Divergence and Evolution", **International journal of biological sciences**, Vol. 3, No. 7, pp. 434-445.

Gupta, R.S. and Mathews, D.W., 2010, "Signature Proteins for the Major Clades of Cyanobacteria", **BMC Evolutionary Biology**, Vol. 10, No., p. 24.

Han, D., Fan, Y. and Hu, Z., 2009, "An Evaluation of Four Phylogenetic Markers in Nostoc: Implications for Cyanobacterial Phylogenetic Studies at the Intrageneric Level", **Current Microbiology**, Vol. 58, No. 2, pp. 170-176.

Hardison, R.C., 2003, "Comparative Genomics", **PLoS biology**, Vol. 1, No. 2, p. E58.

Hess, W.R., 2004, "Genome Analysis of Marine Photosynthetic Microbes and Their Global Role", **Current Opinion in Biotechnology**, Vol. 15, No. 3, pp. 191-198.

Honda, D., Yokota, A. and Sugiyama, J., 1999, "Detection of Seven Major Evolutionary Lineages in Cyanobacteria Based on the 16s Rrna Gene Sequence Analysis with New Sequences of Five Marine Synechococcus Strains", **Journal of Molecular Evolution**, Vol. 48, No. 6, pp. 723-739.

Huson, D.H. and Bryant, D., 2006, "Application of Phylogenetic Networks in Evolutionary Studies", **Molecular Biology and Evolution**, Vol. 23, No. 2, pp. 254-267.

Huynen, M.A., Snel, B. and van Noort, V., 2004, "Comparative Genomics for Reliable Protein-Function Prediction from Genomic Data", **Trends in Genetics**, Vol. 20, No. 8, pp. 340-344.

Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G.M., Richardson, P. and Chisholm, S.W., 2007, "Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus", **PLoS genetics**, Vol. 3, No. 12, p. e231.

Koonin, E.V., 2003, "Comparative Genomics, Minimal Gene-Sets and the Last Universal Common Ancestor", **Nature Review Microbiology**, Vol. 1, No. 2, pp. 127-136.

Koonin, E.V., Aravind, L. and Kondrashov, A.S., 2000, "The Impact of Comparative Genomics on Our Understanding of Evolution", **Cell**, Vol. 101, No. 6, pp. 573-576.

Kunin, V. and Ouzounis, C.A., 2003, "The Balance of Driving Forces During Genome Evolution in Prokaryotes", **Genome Research**, Vol. 13, No. 7, pp. 1589-1594.

Kuzniar, A., van Ham, R.C., Pongor, S. and Leunissen, J.A., 2008, "The Quest for Orthologs: Finding the Corresponding Gene across Genomes", **Trends in Genetics**, Vol. 24, No. 11, pp. 539-551.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S., 2003, "Orthomcl: Identification of Ortholog Groups for Eukaryotic Genomes", **Genome Research**, Vol. 13, No. 9, pp. 2178-2189.

Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F. and Chisholm, S.W., 2004, "Transfer of Photosynthesis Genes to and from Prochlorococcus Viruses", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 101, No. 30, pp. 11013-11018.

Luo, H., Shi, J., Arndt, W., Tang, J. and Friedman, R., 2008, "Gene Order Phylogeny of the Genus Prochlorococcus", **PLoS One**, Vol. 3, No. 12, p. e3837.

Luque, I., Riera-Alberola, M.L., Andujar, A. and Ochoa de Alda, J.A., 2008, "Intraphylum Diversity and Complex Evolution of Cyanobacterial Aminoacyl-Trna Synthetases", **Molecular Biology and Evolution**, Vol. 25, No. 11, pp. 2369-2389.

Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N., Shakhova, V., Grigoriev, I., Lou, Y., Rohksar, D., Lucas, S., Huang, K., Goodstein, D.M., Hawkins, T., Plengvidhya, V., Welker, D., Hughes, J., Goh, Y., Benson, A., Baldwin, K., Lee, J.H., Diaz-Muniz, I., Dosti, B., Smeianov, V., Wechter, W., Barabote, R., Lorca, G., Altermann, E., Barrangou, R., Ganesan, B., Xie, Y., Rawsthorne, H., Tamir, D., Parker, C., Breidt, F., Broadbent, J., Hutkins, R., O'Sullivan, D., Steele, J., Unlu, G., Saier, M., Klaenhammer, T., Richardson, P., Kozyavkin, S., Weimer, B. and Mills, D., 2006, "Comparative Genomics of the Lactic Acid Bacteria", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 103, No. 42, pp. 15611-15616.

Makarova, K.S. and Koonin, E.V., 2007, "Evolutionary Genomics of Lactic Acid Bacteria", **Journal of Bacteriology**, Vol. 189, No. 4, pp. 1199-1208.

Makarova, K.S., Sorokin, A.V., Novichkov, P.S., Wolf, Y.I. and Koonin, E.V., 2007, "Clusters of Orthologous Genes for 41 Archaeal Genomes and Implications for Evolutionary Genomics of Archaea", **Biology Direct**, Vol. 2, No., p. 33.

Marri, P.R., Hao, W. and Golding, G.B., 2006, "Gene Gain and Gene Loss in Streptococcus: Is It Driven by Habitat?", **Molecular Biology and Evolution**, Vol. 23, No. 12, pp. 2379-2391.

Martin, K.A., Siefert, J.L., Yerrapragada, S., Lu, Y., McNeill, T.Z., Moreno, P.A., Weinstock, G.M., Widger, W.R. and Fox, G.E., 2003, "Cyanobacterial Signature Genes", **Photosynthetic Research**, Vol. 75, No. 3, pp. 211-221.

Millard, A.D., Zwirglmaier, K., Downey, M.J., Mann, N.H. and Scanlan, D.J., 2009, "Comparative Genomics of Marine Cyanomyoviruses Reveals the Widespread Occurrence of Synechococcus Host Genes Localized to a Hyperplastic Region: Implications for Mechanisms of Cyanophage Evolution", **Environmental Microbiology**, Vol. 11, No. 9, pp. 2370-2387.

Mirkin, B.G., Fenner, T.I., Galperin, M.Y. and Koonin, E.V., 2003, "Algorithms for Computing Parsimonious Evolutionary Scenarios for Genome Evolution, the Last Universal Common Ancestor and Dominance of Horizontal Gene Transfer in the Evolution of Prokaryotes", **BMC Evolutionary Biology**, Vol. 3, No., p. 2.

Mulkidjanian, A.Y., Koonin, E.V., Makarova, K.S., Mekhedov, S.L., Sorokin, A., Wolf, Y.I., Dufresne, A., Partensky, F., Burd, H., Kaznadzey, D., Haselkorn, R. and Galperin, M.Y., 2006, "The Cyanobacterial Genome Core and the Origin of Photosynthesis", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 103, No. 35, pp. 13126-13131.

Omelchenko, M.V., Wolf, Y.I., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., Daly, M.J., Koonin, E.V. and Makarova, K.S., 2005, "Comparative Genomics of Thermus Thermophilus and Deinococcus Radiodurans: Divergent Routes of Adaptation to Thermophily and Radiation Resistance", **BMC Evolutionary Biolology**, Vol. 5, No., p. 57.

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J.A., Ininbergs, K., Zheng, W.W., Lapidus, A., Lowry, S., Haselkorn, R. and Bergman, B., 2010, "Genome Erosion in a Nitrogen-Fixing Vertically Transmitted Endosymbiotic Multicellular Cyanobacterium", **PLoS One**, Vol. 5, No. 7, p. e11486.

Remm, M., Storm, C.E. and Sonnhammer, E.L., 2001, "Automatic Clustering of Orthologs and in-Paralogs from Pairwise Species Comparisons", **Journal of Molecular Biology**, Vol. 314, No. 5, pp. 1041-1052.

Rogers, M.B., Patron, N.J. and Keeling, P.J., 2007, "Horizontal Transfer of a Eukaryotic Plastid-Targeted Protein Gene to Cyanobacteria", **BMC Biology**, Vol. 5, No., p. 26.

Sandaa, R.A., Clokie, M. and Mann, N.H., 2008, "Photosynthetic Genes in Viral Populations with a Large Genomic Size Range from Norwegian Coastal Waters", **FEMS microbiology ecology**, Vol. 63, No. 1, pp. 2-11.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., Pinter, R.Y., Partensky, F., Koonin, E.V., Wolf, Y.I., Nelson, N. and Beja, O., 2009, "Photosystem I Gene Cassettes Are Present in Marine Virus Genomes", **Nature**, Vol. 461, No. 7261, pp. 258-262.

Shi, T. and Falkowski, P.G., 2008, "Genome Evolution in Cyanobacteria: The Stable Core and the Variable Shell", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 105, No. 7, pp. 2510-2515.

Stal, L.J., 2009, "Is the Distribution of Nitrogen-Fixing Cyanobacteria in the Oceans Related to Temperature?", **Environmental Microbiology**, Vol. 11, No. 7, pp. 1632-1645.

Storm, C.E. and Sonnhammer, E.L., 2002, "Automated Ortholog Inference from Phylogenetic Trees and Calculation of Orthology Reliability", **Bioinformatics**, Vol. 18, No. 1, pp. 92-99.

Swingley, W.D., Blankenship, R.E. and Raymond, J., 2008, "Integrating Markov Clustering and Molecular Phylogenetics to Reconstruct the Cyanobacterial Species Tree from Conserved Protein Families", **Molecular Biology and Evolution**, Vol. 25, No. 4, pp. 643-654.

Tamura, K., Dudley, J., Nei, M. and Kumar, S., 2007, "Mega4: Molecular Evolutionary Genetics Analysis (Mega) Software Version 4.0", **Molecular Biology and Evolution**, Vol. 24, No. 8, pp. 1596-1599.

Tan, L.T., 2007, "Bioactive Natural Products from Marine Cyanobacteria for Drug Discovery", **Phytochemistry**, Vol. 68, No. 7, pp. 954-979.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A., 2003, "The Cog Database: An Updated Version Includes Eukaryotes", **BMC Bioinformatics**, Vol. 4, No., p. 41.

Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V., 2000, "The Cog Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution", **Nucleic Acids Research**, Vol. 28, No. 1, pp. 33-36.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V., 2001, "The Cog Database: New Developments in Phylogenetic Classification of Proteins from Complete Genomes", **Nucleic Acids Research**, Vol. 29, No. 1, pp. 22-28.

Ting, C.S., Rocap, G., King, J. and Chisholm, S.W., 2002, "Cyanobacterial Photosynthesis in the Oceans: The Origins and Significance of Divergent Light-Harvesting Strategies", **Trends in Microbiology**, Vol. 10, No. 3, pp. 134-142.

Tomitani, A., Knoll, A.H., Cavanaugh, C.M. and Ohno, T., 2006, "The Evolutionary Diversification of Cyanobacteria: Molecular-Phylogenetic and Paleontological Perspectives", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 103, No. 14, pp. 5442-5447.

Wall, D.P., Fraser, H.B. and Hirsh, A.E., 2003, "Detecting Putative Orthologs", **Bioinformatics**, Vol. 19, No. 13, pp. 1710-1711.

Wang, Z., Zhu, X.G., Chen, Y., Li, Y., Hou, J. and Liu, L., 2006, "Exploring Photosynthesis Evolution by Comparative Analysis of Metabolic Networks between Chloroplasts and Photosynthetic Bacteria", **BMC Genomics**, Vol. 7, No., p. 100.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L. and Koonin, E.V., 2001, "Genome Trees Constructed Using Five Different Approaches Suggest New Major Bacterial Clades", **BMC Evolutionary Biology**, Vol. 1, No., p. 8.

Xiong, J., 2006, "Photosynthesis: What Color Was Its Origin?", **Genome Biology**, Vol. 7, No. 12, p. 245.

Zauner, S., Lockhart, P., Stoebe-Maier, B., Gilson, P., McFadden, G.I. and Maier, U.G., 2006, "Differential Gene Transfers and Gene Duplications in Primary and Secondary Endosymbioses", **BMC Evolutionary Biolology**, Vol. 6, No., p. 38.

Zehr, J.P., Waterbury, J.B., Turner, P.J., Montoya, J.P., Omoregie, E., Steward, G.F., Hansen, A. and Karl, D.M., 2001, "Unicellular Cyanobacteria Fix N2 in the Subtropical North Pacific Ocean", **Nature**, Vol. 412, No. 6847, pp. 635-638.

# APPENDIX A

Full photosynthesis-related genes in ancestral cyanobacterial genomes

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes. The description of cyanobacterial group according to figure 4.11 relate to their biological property and environmental niche.

| cyanoCOG ID | Gene products | LCCA | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|---|---|
| | **Photosystem I proteins** | | | | | | | | |
| Ortho00513 | photosystem I P700 chlorophyll a apoprotein A1 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00107 | photosystem I P700 chlorophyll a apoprotein A2 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00584 | photosystem I subunit VII | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00304 | photosystem I reaction center subunit II (PsaD) | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00132 | photosystem I reaction center subunit IV | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00604 | Photosystem I reaction center subunit III (PsaF) | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02886 | photosystem I protein PsaK | Y | N | Y | N | Y | Y | Y | N |
| Ortho02619 | photosystem I reaction center subunit VIII | Y | N | N | N | N | N | N | Y |
| Ortho03279 | photosystem I reaction center subunit VIII | N | N | N | N | Y | Y | N | N |
| Ortho11215 | photosystem I reaction center subunit VIII | N | N | N | N | N | N | Y | N |
| Ortho12961 | photosystem I reaction center subunit VIII | N | N | N | N | N | N | N | N |
| Ortho14395 | photosystem I reaction center subunit VIII | N | N | N | N | N | N | N | N |
| Ortho01113 | photosystem I reaction center subunit IX | Y | N | Y | Y | Y | Y | Y | Y |
| Ortho09237 | photosystem I reaction centre subunit IX / PsaJ | N | N | N | N | N | N | N | N |
| Ortho10858 | photosystem I reaction center subunit IX | N | N | N | N | N | N | N | N |
| Ortho01114 | photosystem I reaction center subunit X | Y | N | Y | Y | Y | Y | Y | Y |
| Ortho11135 | photosystem I reaction center protein PsaK | N | N | N | N | N | N | N | N |
| Ortho06714 | photosystem I reaction center subunit X-like protein | N | N | N | Y | N | N | N | N |
| Ortho00109 | photosystem I protein PsaL | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho04693 | photosystem I reaction center subunit XII | N | N | N | N | N | N | N | N |
| Ortho12022 | photosystem I reaction centre subunit XII (PsaM) | N | N | N | N | N | N | N | N |
| Ortho02951 | photosystem I reaction center subunit XII | Y | N | N | Y | Y | Y | N | N |
| Ortho10863 | photosystem I reaction center subunit XII | N | N | N | N | N | N | N | N |
| Ortho05333 | photosystem I 4.8K protein | N | N | N | Y | N | N | N | N |
| Ortho00213 | photosystem I assembly protein Ycf3 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00759 | photosystem I assembly related protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00866 | photosystem I assembly protein Ycf4 | Y | Y | Y | Y | Y | Y | Y | Y |
| | **Photosystem II proteins** | | | | | | | | |
| Ortho00757 | photosystem II 11 kD protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00726 | photosystem II reaction center protein Psb28 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02661 | photosystem II reaction center protein Psb28 | N | N | N | N | N | Y | Y | N |
| Ortho00007 | photosystem II D1 protein PsbA | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho04261 | photosystem II D1 protein PsbA | N | N | N | N | N | N | N | N |
| Ortho05981 | photosystem II reaction centre protein PsbA/D1 | N | N | N | N | N | N | N | N |
| Ortho10813 | photosystem II D1 protein | N | N | N | N | N | N | N | N |
| Ortho00825 | photosystem II chlorophyll-binding protein CP47 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho05161 | Photosystem II protein | N | N | N | N | N | N | N | N |
| Ortho00024 | photosystem II D2 protein PsbD | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00283 | photosystem II reaction center protein H | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho04013 | photosystem II reaction center I protein I | N | N | N | N | N | N | N | N |
| Ortho06561 | photosystem II protein I | N | N | N | N | N | N | N | N |
| Ortho10897 | photosystem II reaction center I protein I | N | N | N | N | N | N | N | N |
| Ortho01370 | photosystem II reaction center protein J | Y | Y | Y | Y | N | N | N | Y |
| Ortho03390 | photosystem II reaction center protein K | N | N | N | N | N | N | N | N |
| Ortho03726 | photosystem II reaction center protein K | N | N | N | N | N | N | N | N |
| Ortho05717 | photosystem II reaction center protein K | N | N | N | N | N | N | N | N |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho11853 | photosystem II reaction center protein K | N | N | N | N | N | N | N | N |
| Ortho04305 | photosystem II reaction center L | N | N | N | N | N | N | N | N |
| Ortho11421 | photosystem II reaction center L | N | N | N | N | N | N | N | N |
| Ortho14145 | photosystem II protein L | N | N | N | N | N | N | N | N |
| Ortho14693 | photosystem II protein PsbL | N | N | N | N | N | N | N | N |
| Ortho01645 | photosystem II protein PsbM | Y | Y | Y | Y | Y | Y | N | Y |
| Ortho04319 | photosystem II protein M PsbM | N | N | N | N | N | N | N | N |
| Ortho11930 | photosystem II reaction center M protein (PsbM) | N | N | N | N | N | N | N | N |
| Ortho01386 | photosystem II protein PsbN | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho05684 | photosystem II reaction centre N prot | N | N | N | N | N | N | N | N |
| Ortho05718 | photosystem II reaction center protein N | N | N | N | N | N | N | N | N |
| Ortho00687 | photosystem II manganese-stabilizing protein PsbO | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00674 | photosystem II PsbP protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01366 | photosystem II protein PsbQ | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02978 | photosystem II protein PsbT | Y | N | N | N | N | N | N | Y |
| Ortho04012 | photosystem II reaction center protein T | N | N | N | N | N | N | N | N |
| Ortho15519 | photosystem II protein PsbT | N | N | N | N | N | N | N | N |
| Ortho01197 | photosystem II complex extrinsic protein precursor U | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho15356 | photosystem II 12 kDa extrinsic protein PsbU | N | N | N | N | N | N | N | N |
| Ortho02418 | photosystem II protein PsbX | Y | N | N | Y | N | N | N | Y |
| Ortho05152 | photosystem II protein X PsbX | N | N | N | N | N | N | N | N |
| Ortho10844 | photosystem II reaction center protein PsbX | N | N | N | N | N | N | N | N |
| Ortho11185 | photosystem II PsbX protein | N | N | N | N | N | N | Y | N |
| Ortho11865 | photosystem II protein X PsbX | N | N | N | N | N | N | N | N |
| Ortho03765 | photosystem II protein Y | Y | N | N | N | N | N | N | Y |
| Ortho04042 | Photosystem II protein PsbY | N | N | N | N | N | N | Y | N |
| Ortho05734 | photosystem II protein Y | N | N | N | N | N | N | N | N |
| Ortho08633 | photosystem II PsbY protein-related protein | N | N | N | N | N | N | N | N |
| Ortho11084 | photosystem II protein PsbY | N | N | N | N | N | N | N | N |
| Ortho12014 | photosystem II protein Y | N | N | N | N | N | N | N | N |
| Ortho14966 | photosystem II protein PsbY | N | N | N | N | N | N | N | N |
| Ortho02428 | photosystem II core protein PsbZ | N | N | N | Y | Y | Y | N | N |
| Ortho03574 | putative photosystem II reaction center protein Z | N | N | N | N | N | N | N | N |
| Ortho03772 | photosystem II reaction center protein Z | Y | N | N | N | N | N | N | Y |
| Ortho04350 | photosystem II reaction center protein PsbZ | N | N | N | N | N | N | N | N |
| Ortho11207 | photosystem II reaction center protein Z | N | N | N | N | N | N | Y | N |

| | **Phycobillisome proteins** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho02685 | phycobilisome degradation family protein | N | N | N | Y | N | N | N | N |
| Ortho00714 | phycobilisome rod-core linker polypeptide cpcG (L-RC 28.5) | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01063 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01127 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01255 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01332 | allophycocyanin beta subunit ApcB | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01353 | phycobilisome LC linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01354 | Phycobilisome protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01389 | Phycobilisome protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01494 | Phycobilisome protein | Y | Y | N | N | N | N | Y | Y |
| Ortho01950 | phycobilisome linker protein | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho02183 | Phycobilisome protein | Y | N | N | N | N | N | N | Y |
| Ortho02632 | Phycobilisome protein | Y | Y | N | N | N | N | Y | Y |
| Ortho03770 | phycobilisome rod-core linker polypeptide (L-RC 28.5) | Y | N | N | N | N | N | N | Y |
| Ortho04021 | Phycobilisome protein | N | N | N | N | N | N | N | Y |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho04283 | phycobilisome linker polypeptide | Y | N | N | N | N | N | N | Y |
| Ortho05299 | phycobilisome degradation protein | N | N | N | N | N | N | N | N |
| Ortho05905 | phycobilisome degradation protein | N | N | N | N | N | N | N | N |
| Ortho06701 | phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho06852 | Phycobilisome degradation protein nblA | N | N | N | N | N | N | N | N |
| Ortho07201 | phycobilisome rod-core linker polypeptide (L-RC 28.5) | N | N | N | N | N | N | N | N |
| Ortho08161 | Phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho08162 | Phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho08634 | phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho09052 | CpcD phycobilisome linker-like | N | N | N | N | N | N | N | N |
| Ortho10703 | phycobilisome linker polypeptide, C-phycoerythrin class II-associated | N | N | N | N | N | N | N | N |
| Ortho11129 | CpcD/allophycocyanin linker domain-containing protein | N | N | N | N | N | N | N | N |
| Ortho11193 | phycobilisome degradation protein NblA | N | N | N | N | N | N | Y | N |
| Ortho12878 | Phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho14339 | CpcD phycobilisome linker-like | N | N | N | N | N | N | N | N |
| Ortho14340 | phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho14341 | phycobilisome protein | N | N | N | N | N | N | N | N |

| | **Chlorophyll-binding proteins** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho00051 | high light inducible protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00065 | high light inducible protein | Y | Y | Y | N | Y | Y | Y | Y |
| Ortho00699 | high light inducible protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01021 | chlorophyll a/b binding light harvesting protein PcbD | Y | N | Y | Y | Y | Y | Y | Y |
| Ortho01511 | high light inducible protein | Y | N | Y | N | N | N | Y | Y |
| Ortho01918 | high light inducible protein hli8 | N | N | N | N | N | N | N | N |
| Ortho01934 | high light inducible protein hli9 | Y | Y | N | N | N | N | Y | Y |
| Ortho02912 | high light inducible protein hli1 | N | N | N | N | N | N | N | N |
| Ortho03191 | putative high light inducible protein | Y | N | N | N | N | N | Y | Y |
| Ortho03202 | high light inducible protein-like protein | N | N | N | N | N | N | N | N |
| Ortho03762 | high light inducible protein-related protein | Y | N | N | N | N | N | N | Y |
| Ortho03990 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho03991 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho03992 | high light inducible protein hli12 | N | N | N | N | N | N | N | N |
| Ortho03999 | high light inducible protein hli6 | N | N | N | N | N | N | N | N |
| Ortho04112 | high light inducible protein | N | N | N | Y | N | N | N | N |
| Ortho04307 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho04309 | high light inducible protein hli13 | N | N | N | N | N | N | N | N |
| Ortho04626 | high light inducible protein-related protein | Y | N | N | N | N | N | N | Y |
| Ortho05669 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho05676 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho07217 | high light inducible protein-like | N | N | N | N | N | N | N | N |
| Ortho07233 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho07325 | chlorophyll a/b binding light harvesting protein PcbE | N | N | N | N | N | N | N | N |
| Ortho07334 | high light inducible protein hli7 | N | N | N | N | N | N | N | N |
| Ortho07335 | chlorophyll a/b binding light harvesting protein PcbH | N | N | N | N | N | N | N | N |
| Ortho07338 | chlorophyll a/b binding light harvesting protein PcbB | N | N | N | N | N | N | N | N |
| Ortho07339 | chlorophyll a/b-binding light-harvesting protein PcbD | N | N | N | N | N | N | N | N |
| Ortho08659 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho10430 | high light inducible protein | N | N | N | N | N | N | N | N |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho10431 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho10439 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho12141 | high light inducible protein | N | N | N | N | N | N | N | N |
| Ortho12164 | chlorophyll a/b binding light harvesting protein PcbF | N | N | N | N | N | N | N | N |
| Ortho12251 | high light inducible protein | N | N | N | N | N | N | N | N |

| | **Chlorophyll biosynthesis enzyme** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho00645 | bacteriochlorophyll a synthase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00231 | light-independent protochlorophyllide reductase subunit N | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00252 | protoporphyrin IX magnesium chelatase subunit ChlD | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00313 | protoporphyrin IX magnesium-chelatase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00437 | Mg-protoporphyrin IX methyl transferase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00590 | light-independent protochlorophyllide reductase subunit B | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00602 | protochlorophyllide reductase iron-sulfur ATP-binding protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00632 | protochlorophyllide oxidoreductase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01111 | magnesium-protoporphyrin IX monomethyl ester cyclase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01563 | Proto-chlorophyllide reductase 57 kD subunit | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho01914 | magnesium-protoporphyrin IX monomethyl ester cyclase | N | N | Y | Y | N | Y | N | N |
| Ortho02178 | putative light-dependent protochlorophyllide oxido-reductase | N | N | N | N | N | N | N | N |
| Ortho02701 | magnesium protoporphyrin IX chelatase, subunit H | N | N | Y | Y | Y | Y | N | N |
| Ortho04934 | protoporphyrinogen oxidase | N | Y | N | N | Y | N | N | N |
| Ortho06187 | 3,8-divinyl protochlorophyllide a 8-vinyl reductase, putative | N | N | N | N | N | N | N | N |
| Ortho09255 | Protoporphyrinogen oxidase-like protein | N | N | N | N | N | N | N | N |
| Ortho09806 | red chlorophyll catabolite reductase | N | N | N | N | N | N | N | N |

| | **Cytochrome b6f complex subunit** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho00369 | apocytochrome f | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00077 | ferredoxin, 2Fe-2S type, PetF1 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00338 | cytochrome b6 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00339 | cytochrome b6-f complex subunit IV | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00355 | cytochrome oxidase assembly | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00755 | cytochrome b6-f complex iron-sulfur subunit | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00801 | cytochrome b559 subunit alpha | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00837 | putative c-type cytochrome biogenesis protein CcdA | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00838 | putative C-type cytochrome biogenesis protein Ccs1 | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01054 | cytochrome b559 subunit beta | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01151 | cytochrome P450 | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho01690 | cytochrome P450 family protein | Y | N | Y | Y | Y | Y | N | Y |
| Ortho01710 | cytochrome b6-f complex subunit PetG | Y | Y | Y | N | N | N | Y | Y |
| Ortho02109 | cytochrome d ubiquinol oxidase subunit I | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02176 | cytochrome b6-f complex subunit PetN | Y | N | N | N | N | N | Y | Y |
| Ortho02454 | cytochrome d ubiquinol oxidase, subunit II | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho03298 | cytochrome b6/f complex subunit VIII | N | N | Y | Y | N | N | N | N |
| Ortho04256 | cytochrome P450 family protein | N | N | N | N | N | N | N | N |
| Ortho04371 | putative cytochrome b6-f complex subunit | N | N | N | N | N | Y | N | N |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho04671 | cytochrome b6-f complex subunit PetM | N | N | N | N | N | N | N | N |
| Ortho04681 | cytochrome b559 subunit beta | Y | N | N | N | N | N | N | N |
| Ortho05136 | cytochrome b(C-terminal)/b6/petD | N | N | N | N | N | N | N | N |
| Ortho05488 | cytochrome P450 family protein | N | N | N | N | N | N | N | N |
| Ortho05508 | cytochrome b6-like protein | N | N | N | Y | N | N | N | N |
| Ortho05606 | cytochrome b6, putative | N | N | N | N | N | N | N | N |
| Ortho06350 | cytochrome oxidase c subunit VIb | N | N | N | N | N | N | N | N |
| Ortho06622 | cytochrome P-450 like protein | N | N | N | N | N | N | N | N |
| Ortho06685 | cytochrome P450 | N | N | N | N | N | N | N | N |
| Ortho07089 | cytochrome b subunit of nitric oxide reductase | N | N | N | N | N | N | N | N |
| Ortho07933 | cytochrome b6-f complex subunit PetM | N | N | N | Y | N | N | N | N |
| Ortho08606 | cytochrome b6f complex subunit PetL | N | N | N | N | N | N | N | N |
| Ortho09115 | cytochrome P450 | N | N | N | N | N | N | N | N |
| Ortho09220 | cytochrome B561 | N | N | N | N | N | N | N | N |
| Ortho09723 | cytochrome P450 | N | N | N | N | N | N | N | N |
| Ortho10360 | cytochrome P450 family protein | N | N | N | N | N | N | N | N |
| Ortho11053 | cytochrome b6-f complex subunit PetM | N | N | N | N | N | N | N | N |
| Ortho11221 | cytochrome b6-f complex subunit PetM | N | N | N | N | N | N | Y | N |
| Ortho11861 | cytochrome b6-f complex subunit PetL | N | N | N | N | N | N | N | N |
| Ortho11998 | cytochrome ubiquinol oxidase | N | N | N | N | N | N | N | N |
| Ortho13116 | cytochrome P450 | N | N | N | N | N | N | N | N |
| Ortho13640 | cytochrome b6/f complex subunit 5 | N | N | N | N | N | N | N | N |
| Ortho14251 | cytochrome b6-f complex subunit PetL | N | N | N | N | N | N | N | N |
| Ortho14580 | cytochrome bd ubiquinol oxidase, subunit II | N | N | N | N | N | N | N | N |

| | **Water-soluble electron carriers** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho01082 | plastocyanin | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00087 | cytochrome c oxidase, subunit I | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00097 | cytochrome c oxidase, subunit II | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00104 | cytochrome c oxidase subunit III | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00115 | cytochrome c6 PetJ | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00579 | cytochrome c assembly protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00845 | cytochrome cM | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01202 | cytochrome c6 (soluble cytochrome f) (cytochrome c553) | Y | N | Y | Y | Y | Y | Y | Y |
| Ortho01228 | cytochrome c, class I | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01613 | cytochrome c oxidase, subunit II | Y | N | N | N | N | N | N | Y |
| Ortho01737 | cytochrome c oxidase, subunit I | Y | N | N | N | N | N | N | Y |
| Ortho01738 | cytochrome c oxidase subunit III | Y | N | N | N | N | N | N | Y |
| Ortho03565 | cytochrome C biogenesis protein transmembrane region protein | Y | N | N | N | N | N | N | Y |
| Ortho04210 | cytochrome c-550 | N | Y | Y | N | Y | N | N | N |
| Ortho05747 | Cytochrome c biogenesis protein, transmembrane region | N | N | N | N | N | N | N | N |
| Ortho09226 | cytochrome c biogenesis protein transmembrane region | N | N | N | N | N | N | N | N |
| Ortho10411 | cytochrome c551 peroxidase | N | N | N | N | N | N | N | N |
| Ortho11261 | cb-type cytochrome c oxidase subunit I | N | N | N | N | N | N | Y | N |
| Ortho14307 | cytochrome c oxidase subunit III | N | N | N | N | N | N | N | N |

| | **Calvin cycle enzymes** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho01860 | phosphoribulokinase | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho01800 | RbcX chaperonin protein | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho01814 | rubisco operon transcriptional regulator RbcR | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho01921 | phosphoribulokinase | Y | N | N | N | N | N | N | Y |
| Ortho04116 | phosphoribulokinase | N | N | N | Y | Y | N | N | N |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | **Regulatory and uncharacterized chloroplast proteins** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho01486 | GUN4 domain protein | N | N | N | Y | Y | N | N | N |
| Ortho00800 | Ycf48-like protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02994 | GUN4-like family protein | N | N | N | N | N | N | N | N |
| Ortho03137 | GUN4 domain protein | N | N | N | Y | N | N | N | N |
| Ortho03712 | GUN4-like family protein | N | N | N | N | N | N | N | N |
| Ortho04494 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho04710 | GUN4 domain protein | N | N | N | N | N | N | N | N |
| Ortho04892 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho05181 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho07063 | GUN4-like family protein | N | N | N | N | N | N | N | N |
| Ortho07455 | GUN4 domain protein | N | N | N | N | N | N | N | N |
| Ortho07466 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho08100 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho08226 | GUN4 domain protein | N | N | N | N | Y | N | N | N |
| Ortho09106 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho09415 | YcfA family protein | N | N | N | N | N | N | N | N |
| Ortho10032 | YcfA-like protein | N | N | N | N | N | N | N | N |
| Ortho10206 | GUN4 domain protein | N | N | N | N | N | N | N | N |
| Ortho13256 | GUN4 domain protein | N | N | N | N | N | N | N | N |
| Ortho13560 | GUN4 domain protein | N | N | N | N | N | N | N | N |
| Ortho13730 | GUN4 domain protein | N | N | N | N | N | N | N | N |

| | **Phycobilisome proteins** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho02685 | phycobilisome degradation family protein | N | N | N | Y | N | N | N | N |
| Ortho00714 | phycobilisome rod-core linker polypeptide cpcG (L-RC 28.5) | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01063 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01127 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01255 | phycobilisome linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01332 | allophycocyanin beta subunit ApcB | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01353 | phycobilisome LC linker polypeptide | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01354 | Phycobilisome protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01389 | Phycobilisome protein | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho01494 | Phycobilisome protein | Y | Y | N | N | N | N | Y | Y |
| Ortho01950 | phycobilisome linker protein | Y | Y | Y | Y | Y | Y | Y | N |
| Ortho02183 | Phycobilisome protein | Y | N | N | N | N | N | N | Y |
| Ortho02632 | Phycobilisome protein | Y | Y | N | N | N | N | Y | Y |
| Ortho03770 | phycobilisome rod-core linker polypeptide (L-RC 28.5) | Y | N | N | N | N | N | N | Y |
| Ortho04021 | Phycobilisome protein | N | N | N | N | N | N | N | Y |
| Ortho04283 | phycobilisome linker polypeptide | Y | N | N | N | N | N | N | Y |
| Ortho05299 | phycobilisome degradation protein | N | N | N | N | N | N | N | N |
| Ortho05905 | phycobilisome degradation protein | N | N | N | N | N | N | N | N |
| Ortho06701 | phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho06852 | Phycobilisome degradation protein nblA | N | N | N | N | N | N | N | N |
| Ortho07201 | phycobilisome rod-core linker polypeptide (L-RC 28.5) | N | N | N | N | N | N | N | N |
| Ortho08161 | Phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho08162 | Phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho08634 | phycobilisome linker polypeptide | N | N | N | N | N | N | N | N |
| Ortho09052 | CpcD phycobilisome linker-like | N | N | N | N | N | N | N | N |
| Ortho10703 | phycobilisome linker polypeptide, C-phycoerythrin class II-associated | N | N | N | N | N | N | N | N |
| Ortho11129 | CpcD/allophycocyanin linker domain-containing protein | N | N | N | N | N | N | N | N |

**Table A.1** All photosynthesis-related genes in ancestral cyanobacterial genomes (contunue).

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Ortho11193 | phycobilisome degradation protein NblA | N | N | N | N | N | N | Y | N |
| Ortho12878 | Phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho14339 | CpcD phycobilisome linker-like | N | N | N | N | N | N | N | N |
| Ortho14340 | phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho14341 | phycobilisome protein | N | N | N | N | N | N | N | N |
| Ortho00176 | phycocyanobilin:ferredoxin oxidoreductase | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho09257 | phycocyanin, beta subunit | N | N | N | N | N | N | N | N |
| Ortho03074 | phycoerythrin-associated linker protein, CpeR | N | N | N | N | N | N | N | N |
| Ortho09256 | phycocyanin, alpha subunit | N | N | N | N | N | N | N | N |
| Ortho01388 | allophycocyanin, beta subunit | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00894 | phycocyanin alpha subunit | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho00893 | phycocyanin beta subunit | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho04320 | phycoerythrin class III gamma chain | N | N | N | N | N | N | N | Y |
| Ortho04284 | C-phycoerythrin class II alpha chain | N | N | N | N | N | N | N | Y |
| Ortho01446 | phycocyanin alpha phycocyanobilin lyase; CpcF | Y | Y | Y | Y | Y | Y | Y | Y |
| Ortho02506 | allophycocyanin alpha subunit ApcA | N | N | Y | N | N | Y | N | N |
| Ortho04643 | phycoerythrin class II gamma chain, linker polypeptide | Y | N | N | N | N | N | N | Y |
| Ortho01580 | phycoerythrin linker protein CpeS-like protein | Y | N | N | N | N | N | N | Y |

The presence or absence of gene coding for proteins of each type in the cyanobacterial ancestral state; Y, gene presence; N, gene absence.

# APPENDIX B

Implemented Python code for

evolutionary scenario algorithm

## B.1) Python module for parsimonious evolutionary scenarios

### B.1.1) Scenario set class

```
#=============================================================#
# Scenario_set class was used to define the gain and lost     #
# event for each node                                         #
#=============================================================#

class scenario_set:
   def __init__(self,Gi,Gn,Li,Ln):
      self.Gi = Gi
      self.Gn = Gn
      self.Li = Li
      self.Ln = Ln
```

### B.1.2) Parsimonious evolutionary scenario class

```
from Bio.Nexus.Trees import Tree
import scenario_set

#=============================================================#
# Define the parsimonious evolutionary scenario class        #
# from Mirkin BG, et al (2003)                                #
#=============================================================#

class par:
   def __init__(self,tree_str,node_has_gene):
      self.tree_str = tree_str
      self.node_has_gene = node_has_gene
      self.tree_obj = Tree(tree_str)
      self.penalty = 1
      self.node_num = []
      for org in node_has_gene:
         node = self.tree_obj.search_taxon(org)
         self.node_num.append(node)

   def __assign_scenarios(self,parent_node):
      child_node = self.tree_obj.node(parent_node).get_succ()
      if (len(child_node) == 2):
         if (self.scenario.has_key(child_node[0])):
            if (self.scenario.has_key(child_node[1])):
               pass
            else:
               self.__assign_scenarios(child_node[1])
         else:
            if (self.scenario.has_key(child_node[1])):
               self.__assign_scenarios(child_node[0])
```

```
            else:
               self.__assign_scenarios(child_node[0])
               self.__assign_scenarios(child_node[1])

      if (child_node == []):
         if set([parent_node]).issubset(self.node_num):
            self.scenario[parent_node] =
scenario_set.scenario_set([],[],[],[])
            self.scenario[parent_node].Gn = [parent_node]
         else:
            self.scenario[parent_node] =
scenario_set.scenario_set([],[],[],[])
            self.scenario[parent_node].Li = [parent_node]

      else:
         if ((self.scenario.has_key(child_node[0])) &
(self.scenario.has_key(child_node[1]))):

#=============================================================#
# Start parsimonious evolutionary algorithm                   #
#=============================================================#

            self.scenario[parent_node] =
scenario_set.scenario_set([],[],[],[])

#=============================================================#
# Inheritance Assumption                                      #
#=============================================================#

            e_lost = len(self.scenario[child_node[0]].Gn) +
len(self.scenario[child_node[0]].Ln) +
len(self.scenario[child_node[1]].Gn) +
len(self.scenario[child_node[1]].Ln) + 1
            e_not_lost = len(self.scenario[child_node[0]].Gi) +
len(self.scenario[child_node[0]].Li) +
len(self.scenario[child_node[1]].Gi) +
len(self.scenario[child_node[1]].Li)
            if (e_lost < e_not_lost):
               self.scenario[parent_node].Gi =
list(set(self.scenario[child_node[0]].Gn).union(set(self.scenari
o[child_node[1]].Gn)))
               self.scenario[parent_node].Li = [parent_node]
            if (e_not_lost <= e_lost):
               self.scenario[parent_node].Gi =
list(set(self.scenario[child_node[0]].Gi).union(set(self.scenari
o[child_node[1]].Gi)))
               self.scenario[parent_node].Li =
list(set(self.scenario[child_node[0]].Li).union(set(self.scenari
o[child_node[1]].Li)))
```

```
#============================================================#
# Non Inheritance Assumption                                 #
#============================================================#

            e_gain = len(self.scenario[child_node[0]].Gi) +
len(self.scenario[child_node[0]].Li) +
len(self.scenario[child_node[1]].Gi) +
len(self.scenario[child_node[1]].Li) + self.penalty
            e_not_gain = len(self.scenario[child_node[0]].Gn) +
len(self.scenario[child_node[0]].Ln) +
len(self.scenario[child_node[1]].Gn) +
len(self.scenario[child_node[1]].Ln)
            if (e_gain < e_not_gain):
               self.scenario[parent_node].Gn = [parent_node]
               self.scenario[parent_node].Ln =
list(set(self.scenario[child_node[0]].Li).union(set(self.scenari
o[child_node[1]].Li)))
            if (e_not_gain <= e_gain):
               self.scenario[parent_node].Gn =
list(set(self.scenario[child_node[0]].Gn).union(set(self.scenari
o[child_node[1]].Gn)))
               self.scenario[parent_node].Ln =
list(set(self.scenario[child_node[0]].Ln).union(set(self.scenari
o[child_node[1]].Ln)))

   def scenarios(self):
      root_node = self.tree_obj.root
      self.scenario = {}
      self.__assign_scenarios(root_node)

   def display_all(self):
      print '#\tGi\tLi\tGn\tLn'
      for node in self.tree_obj.all_ids():
         print str(node) + '\t' + str(self.scenario[node].Gi) +
'\t' + str(self.scenario[node].Li) + '\t' +
str(self.scenario[node].Gn) + '\t' + str(self.scenario[node].Ln)

   def display(self,node):
      print '#\tGi\tLi\tGn\tLn'
      print str(node) + '\t' + str(self.scenario[node].Gi) +
'\t' + str(self.scenario[node].Li) + '\t' +
str(self.scenario[node].Gn) + '\t' + str(self.scenario[node].Ln)

   def get_gain_node(self):
      node_ls =
list(set(self.scenario[self.tree_obj.root].Gn).union(set(self.sc
enario[self.tree_obj.root].Gi)))
      return node_ls
```

```
    def get_lost_node(self):
        node_ls =
list(set(self.scenario[self.tree_obj.root].Ln).union(set(self.sc
enario[self.tree_obj.root].Li)))
        return node_ls
```

## B.2) Python code for running parsimonious evolutionary scenarios

```
#!/usr/bin/python

import os
from Bio.Nexus.Trees import Tree

#=============================================================#
# The script for run the parsimonious evolutionary scenario   #
# algorithm                                                   #
#=============================================================#

curr_dir = os.getcwd()
tree_str = open(curr_dir +
'/ribosomal_protein_tree_root.nwk').read()
tree = Tree(tree_str)
org_ls = [x.strip() for x in open(r''+curr_dir +
'/org_name','r')]
ortho_pattern = open(r'' + curr_dir +
'/output/orthodb/COG_patt.db','r')
ortho_patt = {}

for ortho in ortho_pattern:
    pattern = ortho.split('\t')[1]
    org_patt_ls = []
    index = 0
    for i in str(pattern):
        if (i == '1'):
            org_patt_ls.append(org_ls[index])
        index = index +1
    ortho_patt[ortho.split('\t')[0]] = org_patt_ls

for k,v in ortho_patt.items():
    if (set(v).issubset(tree.get_taxa())):
        pass
    else:
        print k + ' have the node that not contain in the tree'

import par
gain_set = {}
lost_set = {}
for i in tree.all_ids():
```

```
    gain_set[i] = []
    lost_set[i] = []

count = 0
for k,v in sorted(ortho_patt.items()):
    par_scen = par.par(tree_str,v)
    par_scen.scenarios()
    gain_ls = par_scen.get_gain_node()
    lost_ls = par_scen.get_lost_node()
    for i in gain_ls:
        gain_set[i].append(k)
    for i in lost_ls:
        lost_set[i].append(k)
    count = count + 1
    if (count%1000 == 0):
        print '===>> ' + k + ' <<==='

file_out = open(r'' + curr_dir +
'/output/gain_lost_node.out','w')
file_out.write('#gain_list\n')
for k,v in sorted(gain_set.items()):
    file_out.write(str(k) + ':' + ','.join(v) + '\n')
file_out.write('#lost_list\n')
for k,v in sorted(lost_set.items()):
    file_out.write(str(k) + ':' + ','.join(v) + '\n')
file_out.close()
```

# CURRICULUM VITAE (I)

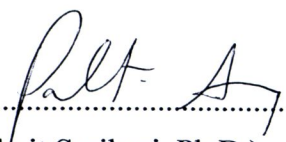
| | |
|---|---|
| **NAME** | Mr. Palang Chotsiri |
| **DATE OF BIRTH** | 1 December 1984 |
| **EDUCATION RECORD** | |
| HIGH SCHOOL | High School Graduation<br>The Laboratory School of Rajabhat Institute Phranakhon Si Ayutthaya, 2003 |
| BACHELOR'S DEGREE | Bachelor of Science (Physics)<br>Mahidol University, 2007 |
| MASTER'S DEGREE | Master of Science (Bioinformatics)<br>King Mongkut's University of Technology Thonburi, 2011 |
| **SCHOLARSHIP** | Full Scholarship, by National Center for Genetic Enginnering and Biotechnology and King Mongkut's University of Technology Thonburi for Master's Degree in Bioinformatics, 2007<br><br>Thailand Full Scholarship for Distinguish Science Student, Ministry of Science and Technology of Thailand for Bachelor's degree in Sciences, 2004 |
| **PUBLICATION** | Chotsiri P., Cheevadhanarak S., Senachak J., Laoteng K., Paithoonrangsarid K., Plengvidhaya V., 2008, "Evolutionary scenarios of cyanobacterial lineage determining by comparative genomics". **International conference on life science 2008 (BioAsia 2008)**, Bangkok, Thailand, (poster presentation)<br><br>Saehuan C., Chotsiri P., Rojanarata T., 2008, "Molecular phylogeny based on benzoylformate decarboxylase of Pseudomonas strains", **ก้าวทันโลกวิทยาศาสตร์ ปีที่ 8(1)**, 78-76.<br><br>Chosiri P., Plengvidhaya V., Senachak J., Paithoonrangsarid K., Laoteng K., Prommeenate P., Cheevadhanarak S., 2009, "Uncovering photosynthesis apparatus and genomic repertoires of cyanobacterial ancestor via phylogenomic analysis". **Thailand Society of Biotechnology (TSB)**. Bangkok, Thailand, (oral presentation) |

Single Nucleotide Polymorphisms (SNPs) and Genome Wide Association Study (GWAS) content pack (II)

Mr. Palang Chotsiri B.Sc. (Physics)

A Thesis Submitted in Partial Fulfillment of the Requirements for
The Degree of Master of Science (Bioinformatics)
School of Bioresources and Technology and School of Information Technology
King Mongkut's University of Technology Thonburi
2009

Thesis Committee



..... Thesis Supervisor (II)
(Pahnit Seriburi, Ph.D.)

| | |
|---|---|
| Thesis Title (II) | Single Nucleotide Polymorphisms (SNPs) and Genome Wide Association Study (GWAS) content pack |
| Thesis Credits (II) | 6 |
| Candidate | Mr. Palang Chotsiri |
| Thesis Supervisor | Dr. Pahnit Seriburi |
| Program | Master of Science |
| Field of Study | Bioinformatics |
| Faculty | School of Bioresources and Technology and School of Information Technology |
| B.E. | 2552 |

Abstract

Single nucleotide polymorphisms (SNPs) are single nucleotide variation in the DNA sequences. Also, genome-wide association study (GWAS) is an examination of genetic variations or SNPs across the human genome to determine the genetic association with the observable traits. The primary use of GWAS for the foreseeable future is likely to be investigation of the biological pathways of disease causation and the normal health and development. An area of substantial future interest for the pharmaceutical industry will be pharmacogenetics and GWAS to identify markers for patient stratification in clinical trials. Nowadays, there are a vast amount of SNPs and GWAS data are stored in the various databases which are public accessible databases. For example, the SNPs database (http://www.ncbi.nlm.nih.gov/projects/SNP/) stores the whole references human SNPs, GWAS catalog (http://www.genome.gov/gwastudies/) stores the data of GWAS and the traits–SNPs associations that have been reported from the prior publications, and international HapMap project (http://www.hapmap.org/) stores and views the SNPs location along the human genome map. However, the GWAS data are still difficult to retrieve and missing several important details. In order to address this issue, the content pack idea is proposed to accumulate, restore and standardize the SNPs and GWAS data.

Keywords: SNPs / GWAS / Content Pack

| | |
|---|---|
| หัวข้อโครงการวิจัย (II) | กล่องข้อมูล ภาวะหลากหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว และ การศึกษาความเชื่อมโยงของทั่วทั้งจีโนม |
| หน่วยกิต (II) | 6 |
| ผู้เขียน | นายพลัง โชติศิริ |
| อาจารย์ที่ปรึกษา | ดร. ผาณิต เสรีบุรี |
| หลักสูตร | วิทยาศาสตรมหาบัณฑิต |
| สาขาวิชา | ชีวสารสนเทศ |
| คณะ | ทรัพยากรชีวภาพและเทคโนโลยี และ เทคโนโลยีสารสนเทศ |
| พ.ศ. | 2552 |


บทคัดย่อ

ภาวะหลากหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว (Single Nucleotide Polymorphisms: SNPs) คือ ความแตกต่างของของนิวคลีโอไทด์เดียวบนลำดับ DNA ในขณะเดียวกัน การศึกษาความเชื่อมโยงของทั่วทั้งจีโนม (Genome-Wide Association Study: GWAS) คือการตรวจสอบความแปรปรวนของสารพันธุกรรม หรือ SNPs ตลอดทั้ง จีโนมของมนุษย์ เพื่อที่จะบ่งชี้ถึงความเชื่อมโยงของสารพันธุกรรมกับคุณลักษณะที่สามารถสังเกตได้ การใช้การศึกษาความเชื่อมโยงของทั่วทั้งจีโนมเบื้องต้นเพื่อทำนายลักษณะที่จะเกิดขึ้นในอนาคต โดยการใช้การตรวจสอบจากวิถีทางชีววิทยา ของสาเหตุของการเกิดโรคเบื้องต้น และ สุขภาพและพัฒนาการปกติ สำหรับในอนาคตที่น่าสนใจของอุสาหกรรมทางด้านเภสัชกรรม คือการใช้ การศึกษาความเชื่อมโยงของทั้งจีโนม เพื่อที่จะ เป็นเครื่องหมายบ่งชี้ สำหรับการแบ่งผู้ป่วยในช่วงการทดลองทางคลีนิค ณ ปัจจุบันนี้ ข้อมูลของภาวะหลายหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว และ การศึกษาความเชื่อมโยงของทั่วทั้งจีโนม ถูกเก็บเปิดเผยข้อมูลบนระบบเครือข่ายทางคอมพิวเตอร์อย่างหลากหลาย ตัวอย่างเช่น ฐานข้อมูลของ ภาวะหลายหลายรูปแบบบนนิวคลีโทไทด์เดี่ยว (SNPs databases: http://www.ncbi.nlm.nih.gov/projects/SNP/) ได้เก็บข้อมูลอ้างอิงทั้งหมดของภาวะหลายหลายรูปแบบบนนิวคลีโทไทด์เดี่ยว ในมนุษย์, รายการการศึกษาความเชื่อมโยงของทั้งจีโนม (GWAS catalog : http://www.genome.gov/gwastudies/) ได้เก็บข้อมูลของ การศึกษาความเชื่อมโยงของทั้งจีโนม และ ภาวะหลายหลายรูปแบบบนนิวคลีโทไทด์เดี่ยวที่สัมพันธ์กับคุณลักษณะที่สังเกตได้ และมีรายงานจากการตีพิมพ์บนวารสารทางวิชาการ และ โครงการ HapMap ระหว่างชาติ (http://www.hapmap.org/) ได้เก็บข้อมูล ของ ตำแหน่งของภาวะหลายหลายรูปแบบบนนิวคลีโทไทด์เดี่ยวในจีโนมมนุษย์ แต่อย่างไรก็ตาม รายละเอียดของ ข้อมูลของการศึกษาความเชื่อมโยงของทั้งจีโนม ยังคงขาดหายไป เพื่อที่จะแก้ปัญหานี้ แนวความคิดเรื่อง กล่องข้อมูล เก็บสะสม ฟื้นฟู และ ทำให้เป็นมาตรฐานเดียวกันของภาวะหลากหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว บนจีโนมมนุษย์ และ ข้อมูลของ การศึกษาความเชื่อมโยงทั่วทั้งจีโนม ได้ถูกนำเสนอขึ้น


คำสำคัญ: ภาวะหลากหลายรูปแบบบนนิวคลีโอไทด์เดี่ยว / การศึกษาความเชื่อมโยงทั่วทั้งจีโนม

# ACKNOWLEDGEMENT (II)

I would like to thank my supervisor, Dr. Pahnit Seriburi, who always supports me with her professional and valuable guidance. Her suggestion and ceaseless contribution were most rewarding and informative to this thesis. I would like thank Dr. Sirimon O-Charoen, Senior Research Scientist at Torrey Path Inc., for their kindness and all the helpful supervision. I would like to thank to Peter Dresslar, President and CEO of Torrey Path Inc., who give me the great opportunity to intern with the world leader company, Torrey Path Inc.

I would like to express my appreciation to my co-worker at Torrey Path Inc., Punnarai Wuthipanyarattanakun, and Poranate Klanrit for their helpful and valuable comments throughout this work.

I am appreciative to all my lecturers in the Bioinformatics program at King Mongkut's University of Technology Thonburi (KMUTT) for all given knowledge and opportunities. I would like to express my gratitude to the program for allowing me to carry out my master study. Furthermore, I would like to convey my deepest gratitude to both National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand and KMUTT for the full scholarship that allows me to finish my master degree.

# CONTENTS (II)

## LIST OF TABLES (II)

# LIST OF FIGURES (II)

# LIST OF ABBREVIATION (II)

| | | |
|---|---|---|
| DNA | = | Deoxyribonucleic acid |
| GWAS | = | Genome wide association study |
| LD | = | Linkage disequilibrium |
| MAF | = | Minor allele frequency |
| SNPs | = | Single nucleotide polymorphisms |

# CHAPTER 1 INTRODUCTION (II)

## 1.1 Background and rationale

Single Nucleotide Polymorphisms (SNPs) are single nucleotide variation in the DNA sequences. The genetic variations of SNPs relate to diseases or health–related traits. Nearby 12 million unique human SNPs have been assigned a reference SNPs number in the National Center of Biotechnology Information's dbSNPs database and characterized to specific alleles (Pearson and Manolio, 2008). Genome-Wide Association Study (GWAS) is an examination of genetic variant or SNPs across the human genome to determine the genetic association with the observable traits.

The primary usage of GWAS for the foreseeable future is likely to be investigation of the biological pathways of disease causation and the normal health and development. An area of substantial future interest for the pharmaceutical industry will be pharmacogenetics GWAS to identify markers for patient stratification in clinical trials. Comprehensive pharmacogenetic information will, in turn, facilitate the practice of personalized medicine. Pharmacogenetic GWAS and early adoption of personalized therapy are likely to be used in the selection of expensive or chronic medications in life threatening conditions or where the therapeutic index is narrow or adverse event concerns are high, such as cancer chemotherapy (Kingsmore, *et al.*, 2008).

Nowadays, there are a lot of SNPs and GWAS data are stored in the various public online databases. For example, the SNPs databases (http://www.ncbi.nlm.nih.gov/projects/SNP/) store the whole references human SNPs, GWAS catalogs (http://www.genome.gov/gwastudies/) store the data of GWAS and the traits–SNPs association that reported from the publication, and international HapMap project (http://www.hapmap.org/) store and view the SNPs location on the human genome. However, the GWAS data are still difficult to retrieve and missing a lot of details. To address this issue, the content pack idea is proposed to accumulate, restore and standardized the SNPs and GWAS data.

The SNPs and GWAS data are rapidly generated, but it not stored in the same place and used the same standard. Also, it is difficult to access or accumulate all information from the online databases and publication. Here, the content pack of SNPs and GWAS data will be constructed for reorganizing and visualizing these data into the same platforms.

## 1.2 Objectives

1) To create the platform for restoring the SNPs and GWAS data
2) To reconstruct the SNPs and GWAS content pack

## 1.3 Scope of work

In order to archive the objective, the scope of work would be accomplished according to these following steps:

1) The platform of the SNPs and GWAS content pack has created by reading the literature in this field. Also, the core content pack was created by summarization of the shared property across studies.
2) The primary SNPs and GWAS content pack was done by annotating some study in a spreadsheet.

## 1.4 Expected outputs

1) The new ideas and methods for using the research literature to create the SNPs and GWAS content pack
2) The primary platform for creating the SNPs and GWAS content pack
3) The primary database of SNPs and GWAS content pack

# CHAPTER 2 LITERATURE REVIEWS (II)

## 2.1 Human genetic variants

Human genetics variants are typically referred to as either common or rare, to denote the frequency of the minor allele in the human population. Common variants are synonymous with polymorphisms, defined as genetics variants with minor allele frequency (MAF) of at least one percent in the population, whereas rare variants have MAF of less than 1%. Genetics variants are also discussed in terms of their nucleotide composition. The variation in the human genome can be divided into two difference nucleotide classes: single nucleotide variants and structural variants, as shown in the figure 2.1 and 2.2 (Frazer, *et al.*, 2009).

### 2.1.1 Single nucleotide variants

SNPs are the most prevalent class of genetic variation among individuals. On the basis of survey sequencing results it has been estimated that the human genome contains at least 11 million SNPs, with ~7 million of these occurring with a MAF of over 5% and the remaining having MAFs between 1 and 5%. Analysis of the four fully sequenced individual genomes suggests that these original estimates are fairly accurate and that most SNPs have been identified and information about them deposit in the Single Nucleotide Polymorphisms databases (dbSNPs). The alleles of SNPs located in the same genomics interval are often correlated with the one another. This correlation structure, or linkage disequilibrium (LD), varies in a complex and unpredictable manner across the genome and between difference populations.

### 2.1.2 Structural variants

Structural variation, broadly defined, refers to all base pairs that differ between individuals and that are not single nucleotide variants. Such variation includes insertion-deletions (in-dels), the block substitutions, inversion of DNA sequences and copy number differences, as illustrated in the figure 2.1. Compared with single nucleotide variants, the technological ability to detect structural variants in the human genome has only recently emerged (Eichler, *et al.*, 2007). Hence our understanding the locations and frequencies of structural variants, and our ability to assay their association with the complex traits, is still maturing. Analysis of the four fully sequenced human genomes combined with the targeted sequencing of structural variants greater than 8kb in length in eight human genomes has provided tremendous insight. These studies suggest that structural variation accounts for at least 20% of all genetics variants in humans and underlies greater than 70% of the variant bases. Altogether, for any given individual, structural variants constitute between 9 and 25 Mb of genome (~0.5 to 1%), underscoring the important roles of this class of variation in genome evolution and in human health and disease.

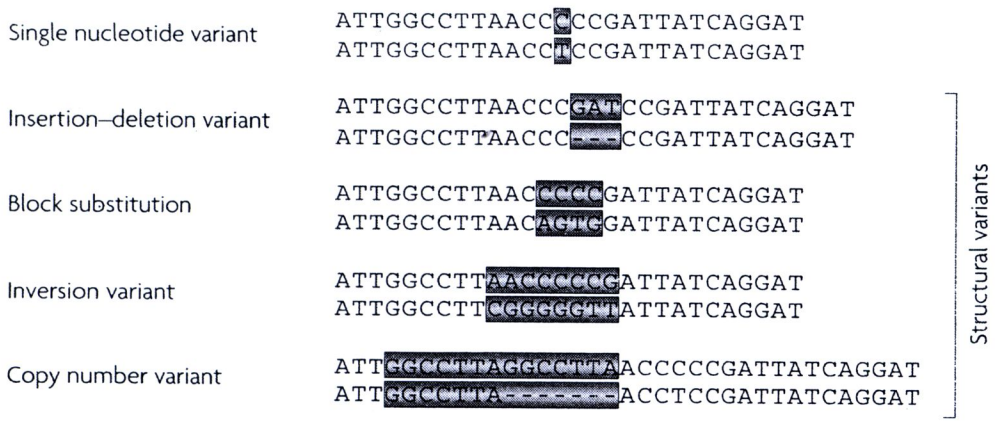



**Figure 2.1** Classes of human genetics variants. Single nucleotide variant are DNA sequence variations in which a single nucleotide (A, C, T and G) is alters. Insertion-deletion variants (in-dels) occur when one or more base pairs are present in some genome and absent in others. Block substitutions describe cases in which a string of adjacent nucleotides varies between two genomes. An inversion variant is one in which the order of the base pairs is reversed in a defined section of a chromosome. Copy number variants occur when identical or nearly identical sequences are repeated in some chromosomes but not others. (Frazer, *et al.*, 2009).

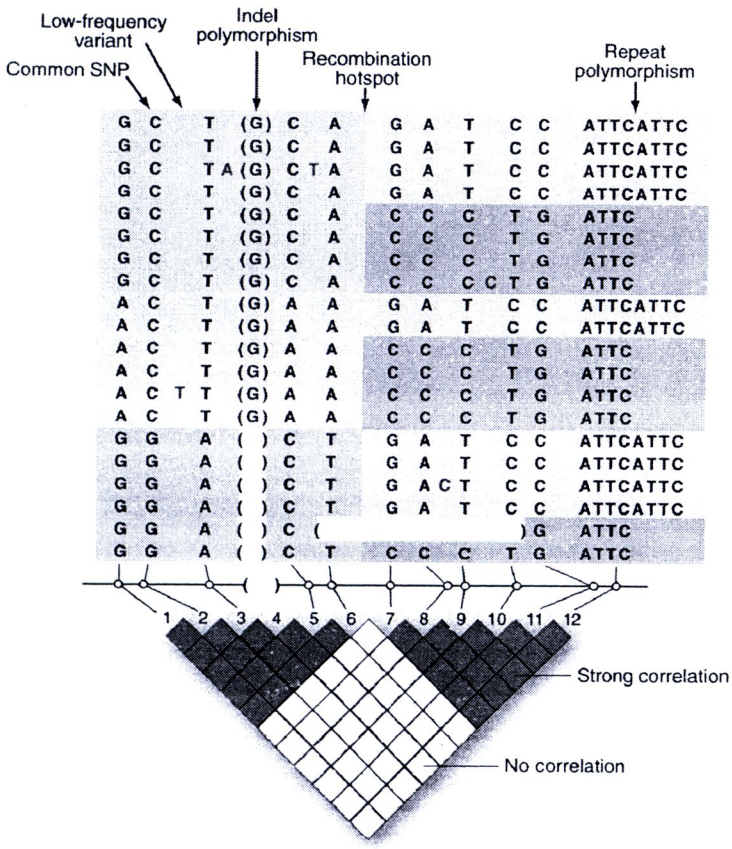



**Figure 2.2** An example common and rare genetic variation in 10 individuals, such as common SNP, low-frequency variant, in-del polymorphism, recombination hotspot, and repeat polymorphism. (Altshuler, *et al.*, 2008)

### 2.1.3 Contribution of variation to phenotypes

In human, hundreds of complex phenotypic traits determine how we behave, and our propensity to develop certain diseases. Each complex phenotype is governed by a combination of inherited factors, which are largely believed to be genetics variants, and environmental influences. Full sequencing of human genomes has shown that in any given individual there are, on average, ~4 million genetics variants encompassing ~12 Mb of sequence. The challenge is to determine which of these variants underlies of its responsible for the inherited components of phenotypes. Over the last decade or so the human genetics field has debate the common disease–common variants hypothesis, which posits that common complex traits are largely due to common variants with small to modest effect. The opposite theory, the rare variants hypothesis, posits that common complex traits are the summation of low-frequency, high-penetration variants (Bodmer and Bonilla, 2008). Overall the field it is making earnest attempts to determine the relative importance of common and rare variants in common complex phenotypic traits.

## 2.2 Genome wide association studies (GWAS)

### 2.2.1 Linking common genetics variants to common complex traits

Concurrent with the efforts of the scientific community to dissect the human genome into linkage disequilibrium (LD) block were extraordinary technological advances in assaying SNPs. From 1997 to 2007, technological advances moved the field from one SNP at a time to assessment of a million SNPs per individual. These two fronts of progress-one on the empirical determination of the LD structure of SNPs across the genome and the other a new-found capacity to perform ultra-high-throughput genotyping- set the foundation for a veritable avalanche of discoveries of common traits and diseases through GWAS. The stage of genome wide association study is illustrated in the figure 2.3.

There are several excellence reviews of GWAS designs and analysis that discuss selection of cases and controls, and statistical analysis – including dealing with population stratification and replication (Pearson and Manolio, 2008). The typical GWAS has 4 parts: (1) selection of a large number of individuals with the disease or trait of interest and a suitable comparison group; (2) DNA isolation, genotyping, and data review to ensure high genotyping quality; (3) statistical tests for associations between SNPs passing quality thresholds and the disease/trait; and (4) replication of identified associations in an independent population sample or examination of functional implications experimentally.

Most of the roughly 100 GWAS studies published by the end of 2007 were designed to identify SNPs associated with common diseases. However, the technique can also be used to identify genetic variants related to quantitative traits such as height or electrocardiographic quantitative traits interval, and to rank the relative importance of previously identified susceptibility genes.

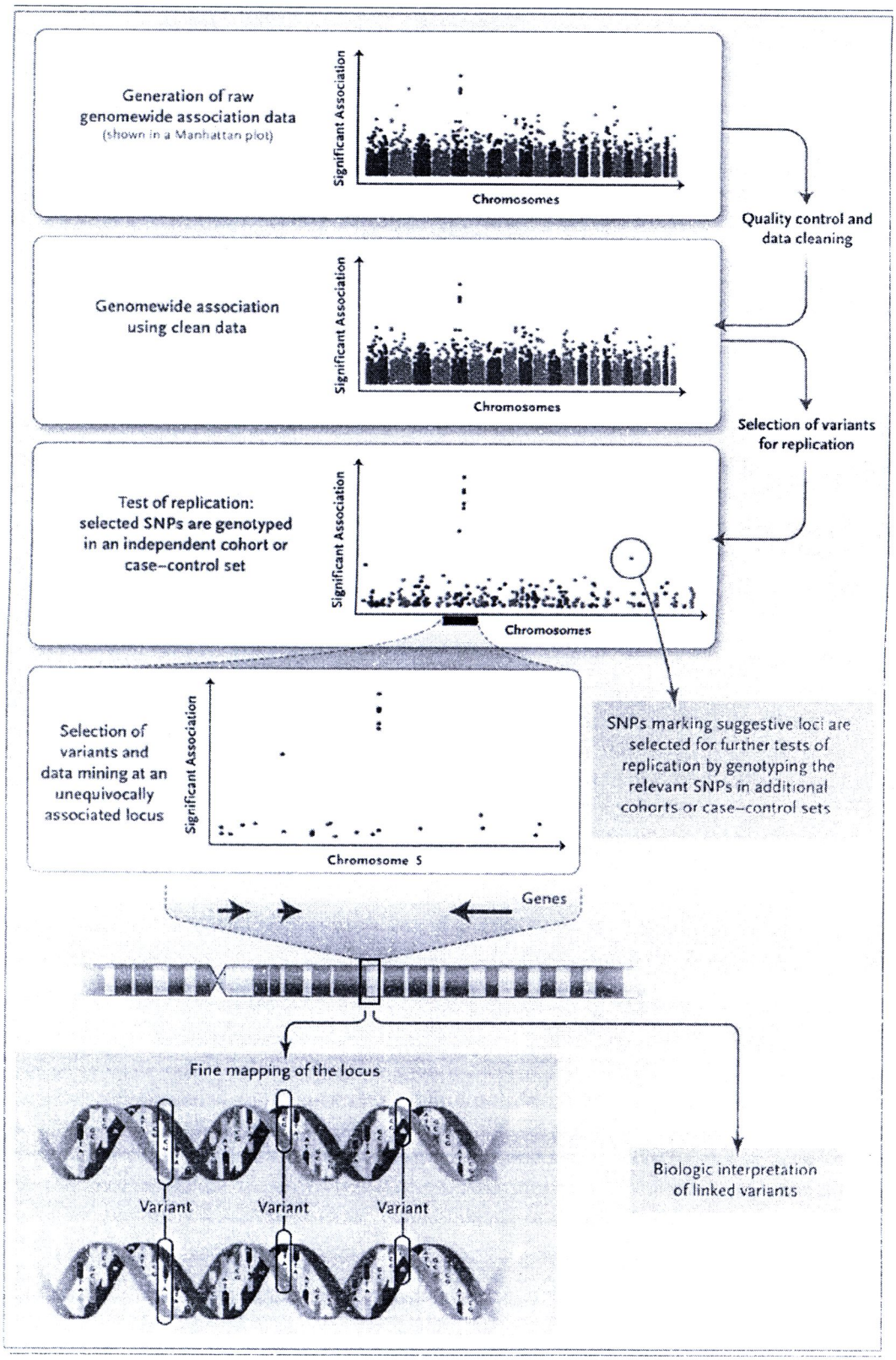



**Figure 2.3** Stage of a genome wide association study (Hardy and Singleton, 2009)

### 2.2.2 Study designs used in GWAS

By far the most frequently used GWAS design to date has been the case-control design, in which allele frequencies in patients with the disease of interest are compared to those in a disease-free comparison group. These studies are often easier and less expensive to conduct than studies using other designs, especially if sufficient numbers of case and control participants can be assembled rapidly. This design also carries the most assumptions, which if not met, can lead to substantial biases and spurious associations, as describe in Table 2.1. The most important of these biases involve the selected, often unrepresentative nature of the study case participants, who are typically sampled from clinical sources and thus may not include fatal, mild, or silent cases not coming to clinical attention; and the lack of comparability of case and control participants, who may differ in important ways that could be related both to genetic risk factors and to disease outcomes (Manolio, *et al.*, 2006).
If well-established principles of epidemiologic design are followed, case-control studies can produce valid results that, especially for rare diseases, may not be obtainable in any other way. However, genetic association studies using case-control methodologies have often not always adhered to these principles. The often sharply abbreviated descriptions of case and control participants and lack of comparison of key characteristics in genome wide association reports can make evaluation of potential biases and replication of findings quite difficult.

The trio design includes the affected case participant and both of his or her parents. Phenotypic assessment (classification of affected status) is performed only in the offspring and only affected offspring are included, but genotyping is performed in all 3 trio members. The frequency with which an allele is transmitted to an affected offspring from heterozygous parents is then estimated. Under the null hypothesis of no association with disease, the transmission frequency for each allele of a given SNP will be 50%, but alleles associated with the disease will be transmitted in excess to the affected case individual. Because the trios design studies allele transmission from parents to offspring, it is not susceptible to population stratification, or genetic differences between case and control participants unrelated to disease but due to sampling them from populations of different ancestry. A significant challenge of the trio design in GWAS is its sensitivity to even small degrees of genotyping error, which can distort transmission proportions between parents and offspring, especially for uncommon alleles. Therefore, standards for genotyping quality in trio studies may need to be more stringent than for other designs.

Cohort studies involve collecting extensive baseline information in a large number of individuals who are then observed to assess the incidence of disease in subgroups defined by genetic variants. Although cohort studies are typically more expensive and take longer to conduct than case-control studies, they often include study participants who are more representative than clinical series of the population from which they are drawn, and they typically include a vast array of health-related characteristics and exposures for which genetic associations can be sought. For these reasons, genome-wide genotyping has recently been added to cohort studies such as the Framingham Heart Study (Cupples, *et al.*, 2007) and the Women's Health Study (Ridker, *et al.*, 2008).

**Table 2.1** Study designs used in GWAS (Pearson and Manolio, 2008).

| | Case-Control | Cohort | Trio |
|---|---|---|---|
| **Assumptions** | • Case and control participants are drawn from the same population<br>• Case participants are representative of all cases of the disease, or limitations on diagnostic specificity and representativeness are clearly specified<br>• Genomic are epidemiology data are collected similarly in cases and controls<br>• Differences in allele frequencies relate to the outcome of interest rather than differences in background population between cases and controls | • Participants under study are more representative of the population from which they are drawn<br>• Diseases and traits are ascertained similarly in individuals with and without gene variant | • Disease-related alleles are transmitted in excess of 50% to affected offspring from heterozygous parents |
| **Advantages** | • Short time frame<br>• Large number of case and control participants can be assembled<br>• Optimal epidemiologic design for studying rare disease | • Cases are incident (developing during observation) and free of survival bias<br>• Direct measure of risk<br>• Fewer biases than case-control studies<br>• Continuum of health-related measures available in population samples not selected for presence of disease | • Control of population structure; immume to population stratification<br>• Allows checks for Mendellian inheritance patterns in genotyping quality control<br>• Logistically simpler for studies of children's condition<br>• Does not require phenotyping of parent |
| **Disadvantages** | • Prone to a number of biases including population stratification<br>• Cases usually prevalent cases, may exclude fatal or short episodes, or mild or silent cases<br>• Overestimate relative risk for common diseases | • Large sample size needed for genotyping if incidence is low<br>• Expensive and lengthy follow-up<br>• Existing consent may be insufficient for GWA genotyping or data sharing<br>• Requires variation in trait being studied<br>• Poorly suited for studying rare diseases | • May be difficult to assemble both parents and offspring, especially in disorders with older ages of onset<br>• Highly sensitive to genotyping error |

GWAS published to date have used various commercials genotyping platforms containing approximately 300,000 to 500,000 common SNPs to detect differences in allele frequencies between cases and controls. Such studies are hypothesis-free, as there is no bias or presumptive list of candidate genes that are being tested. However, the term 'genome-wide' is a misnomer, because approximately 20% of common SNPs are only partially tagged or not tagged at all, and rare variations are generally not tagged. For over 80 phenotypes—including diseases and biological measurements—GWAS have provided remarkably compelling statistical associations for a total of over 300 difference loci in the human genome. The results have been reported on almost weekly basis from April 2007, with over 220 studies reported to date. Almost all disease categories have been addressed, including cardiovascular, neurodegenerative, neuropsychiatric, metabolic, autoimmune and musculoskeletal diseases, and several types of cancer.

### 2.2.3 Enhanced understanding of human diseases

The most impressive outcome of this knowledge base, which connects genomic intervals with complex traits, is a new understanding of the molecular underpinnings and pathways of many diseases. Notably, most of the genes or genomic loci that have been identified through GWAS have not previous been known to be related to the complex trait under investigation. For a substantial number of common diseases the newly identified pathways suggest that molecular sub-phenotypes may exists; that is, although a number of difference pathways might potentially be involved in the development of a particular diseases when all cases are considered, in any individual with the disease only one or a subset of these pathways might be involved. For example, the genetic propensity to develop type 2 diabetes (T2D) seem to involve genes in several different pathways that affect pancreatic $\beta$-cell formation and function, as well as pathways affecting fasting glucose levels and obesity (Frayling, 2007), as illustrated in figure 2.4. Likewise, many of the loci associated with multiple sclerosis involve immune function – including the interleukin receptor genes *IL2RA* and *IL7RA*, and the *HLA-DRA* locus – but a gene encoding a protein involved in axonal function, kinesin family member 1B (*KIF1B*), is also associated with the disease. Clinicians previously considered these conditions as simple phenotypes, with all patients with diagnosis having the same underlying biological disorder.

Surprisingly, there have been several instances in which one genomic interval have been associated with two or more seemingly distinct diseases. This convergence of genes associated with multiple diseases has led to the concepts of the 'diseaseome' , which maps a network of how different genes and pathways connect to various diseases, as illustrated in figure 2.5. Examples include different inter-leukin receptor genes that are associated with Crohn's disease, multiple sclerosis, systemic lupus erythematous and rheumatoid arthritis. Such diseases had already been thought of as sharing a common immune-mediated etiology, but now there is discrete evidence for a common genetic underpinning. Another example is the common SNP on chromosome 9p21 that is associated with three vascular phenotypes — myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Such conditions would not previously have been thought to have a common pathogenic thread. The recent exceptional advances in associating genes with many diseases have led some to suggest that the textbooks of medicine need to rewritten to account for our enhanced understanding of the interconnectivity of the molecular basis underlying distinct diseases.

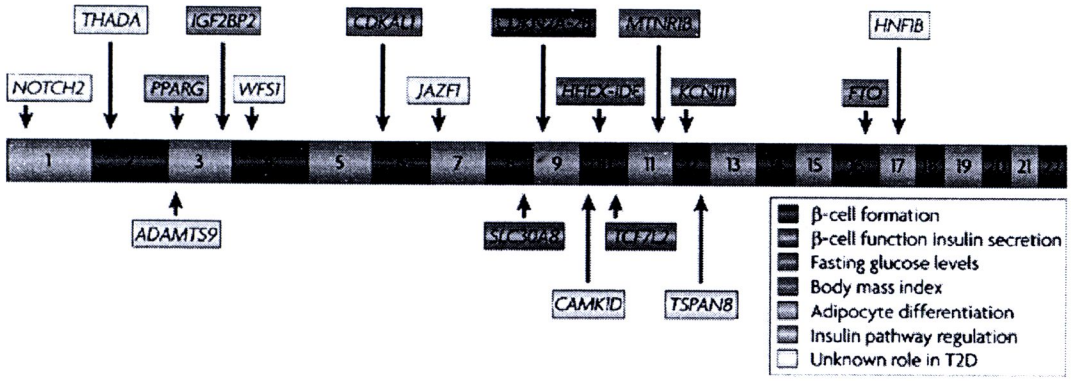



**Figure 2.4** Insights into the genetics basis of type 2diabetes (T2D) (Frazer, *et al.*, 2009).

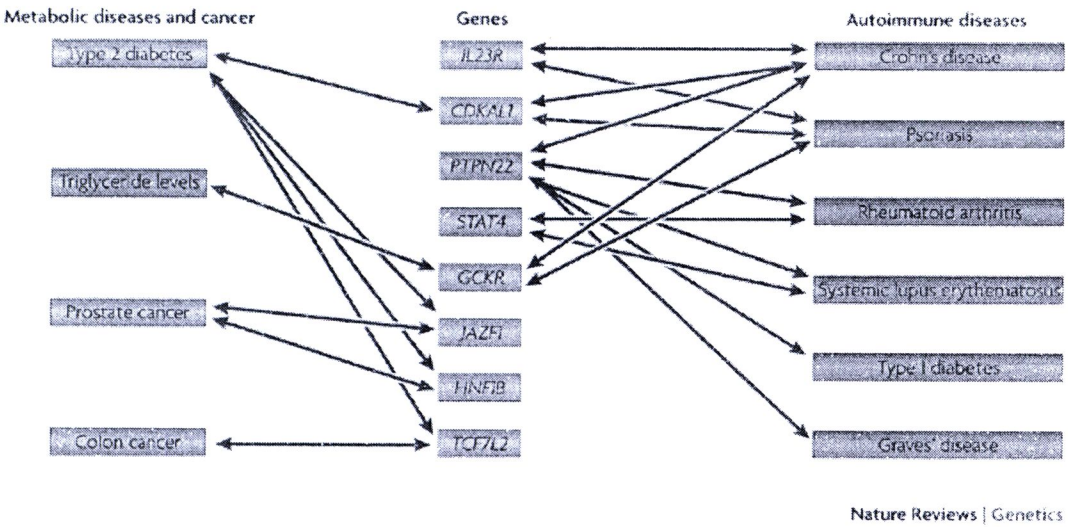



**Figure 2.5** Overlap of genetic risk factor loci of common diseases (Frazer, *et al.*, 2009).

### 2.3.4 Limitations of GWAS in identifying causative variants

Despite this exceptional progress, there are substantial limitations to the GWAS approach. Although statistically compelling associations have been identified, there is an enormous gap in the ability to provide the biological explanation for why a genomic interval tracks with a complex trait. For the most part, all we know is that a tag SNP for an LD bin is statistically associated with a trait, but we have no idea of the precise variants in the bin that have a causal role in contributing to variation in the trait. It is important to emphasize that tag SNPs are in LD not only with other SNPs but also with common structural variants, the majority of which have not yet been identified. The best way to move from a statistical association to knowledge of the causative variant is unclear. In most cases it will be straightforward to identify causative variants that are in LD with a tagging SNP and that are located in exons that truncate or otherwise alter the gene product. However, the causative variants underlying GWAS associations are likely to be regulatory rather than coding. For instance, many of the associations so far are not even localized to intervals that include a gene. For example, the variant at 9p21 that associates with myocardial infarction is 150 kb from the nearest gene, and for the variants on 8q24 that are associated with susceptibility to multiple solid tumors this distance is 300 kb. Experiments are being conducted that simultaneously assay global gene expression and genome-wide variation in a large number of individuals to map genetic factors underlying differences in expression levels. These data sets may be valuable tools for identifying the causative variants and biological bases for many loci associated with a complex trait through genome wide association studies.

# CHAPTER 3 MATERIALS AND METHODS (II)

## 3.1 Materials

### 3.1.1 Databases

The online databases are the major material for constructing the SNPs and GWAS content pack. The following is the databases for this study.

1) PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is the main databases for all of the referenced literature.
2) SNPs database (http:// www.ncbi.nlm.nih.gov/projects/SNP/) are the referenced SNPs information of human.
3) dbGaP (http:// www.ncbi.nlm.nih.gov/gap) is the genotypes and phenotypes databases.
4) International HapMap Project (http:// www.hapmap.org/) is the haplotype databases.
5) Drug Bank (http:// www.drugbank.ca/) is the online databases that combine the drug detail data with a comprehensive drug target.

### 3.1.2 Computer resources

1) CPU: Intel(R) Core(TM) 2 Duo CPU @ 2.20GHz
2) Hard disk: 250 GB
3) Memory: 2 GB

## 3.2 Methods

The methods for constructing the SNPs and GWAS content pack as follows:

1) For finding the main concept and the major component of the GWAS research, the literature review of the previous researches in this field is needed. When the articles from various studies about the SNPs and GWAS are compared, the main idea about the genome wide association study was retrieved. Moreover, the major skeletons and the minor details from each study can be separated.
2) For finding the share components on GWAS research, the major skeleton and the minor detail form each study was separated. For the major skeletons of the genome wide association study, the essential and shared components from each study was included, such as the traits or diseases for each study, the associated SNPs, the statistical test or the experimental design. For the minor details component, the additional component for each study was included, such as the disease annotation, the sample annotation, or the population details.
3) For constructing the SNPs and GWAS content pack, the essential and additional components from the previous steps are summarized and evaluated in order to investigate the standardized fields of SNPs and GWAS content pack.
4) Finally, for testing the constructed SNPs and GWAS content pack, the literature annotation was performed. Few SNPs and GWAS articles were added into the constructed SNPs and GWAS content pack. Moreover, the correlation study across study was performed by mapping the genotypes to phenotypes in order to detect the emerging characteristic of SNPs and GWAS content pack.

# CHAPTER 4 RESULTS AND DISCUSSIONS (II)

## 4.1 Data linking

The data linking of the SNPs and the fields that are associated with the SNPs is showed in the figure 4.1. The SNPs are referenced by location on the chromosome and the allele. On the other hand, the SNPs were defined by its genotypes, also the haplotype for each population groups. For the people that have the difference SNPs, difference genotypes, the gene expression or metabolism are different. The SNPs are associated with the reaction of the body to the drug, for example, the adverse event reaction, the hypersensitivity, or the difference drug metabolism. For the possibility of the occurring of each complex disease, the risk factor is indicated by the SNPs. The online databases and the previous studies were used to fulfill the SNPs data linking. For example, the SNPs databases are provided the reference SNPs information in the area of location on the chromosome, genotype, gene, allele frequency, and the allele frequency for each population groups. The GWAS literatures that are deposit on PubMed are provided the main reference information about the risk factor for each disease, or the drug reaction that are affected by the SNPs.

## 4.2 Structure of SNPs and GWAS content pack

The structure component of SNPs and GWAS content pack can separate into essential and additional component. The essential component of SNPs and GWAS content pack is described by a literature information and SNPs information. The literature information was explained by the publication information, such as, authors, a publication date, journal information, and PubMed information. The study information was explained by the sample information (population information), experimental information, and SNPs information. The additional component of SNPs and GWAS content pack is described by diseases or treats information. The disease information was explained by the type of disease and ontology of diseases and traits. For annotation the literature into the SNPs and GWAS content pack, the annotation fields was separate into 4 parts as following:

1) **General information annotation**
   This part describe the general information about the literature and the study, such as, PubMed ID, authors, publication date, journal, study title, study types, and diseases or traits annotation. The diseases and traits annotation were explained by diseases and traits information, diseases definition, and a disease stage.

2) **Sample annotation**
   The sample annotation describes the sample that used in each study. The sample annotation is composed of an initial sample size, an initial sample detail, a replication sample size and a replication sample detail.

3) **Experimental annotation**
   The experimental annotation used to describe the experimental design of each study. The experimental annotation was explain by a study scope, genotyping methods, genotyping platforms, a quality control cut off, SNPs that passing quality control, and statistical analysis methods.

4) **SNPs annotation**
The SNPs annotation are including by the associated trait, specific traits, SNPs ID, a reference allele, a risk allele, risk allele frequency, chromosomal position, region, SNP type, reporter gene, p-value, log[p-value], odd ratio, 95% confidential interval, and the copy number variation information.

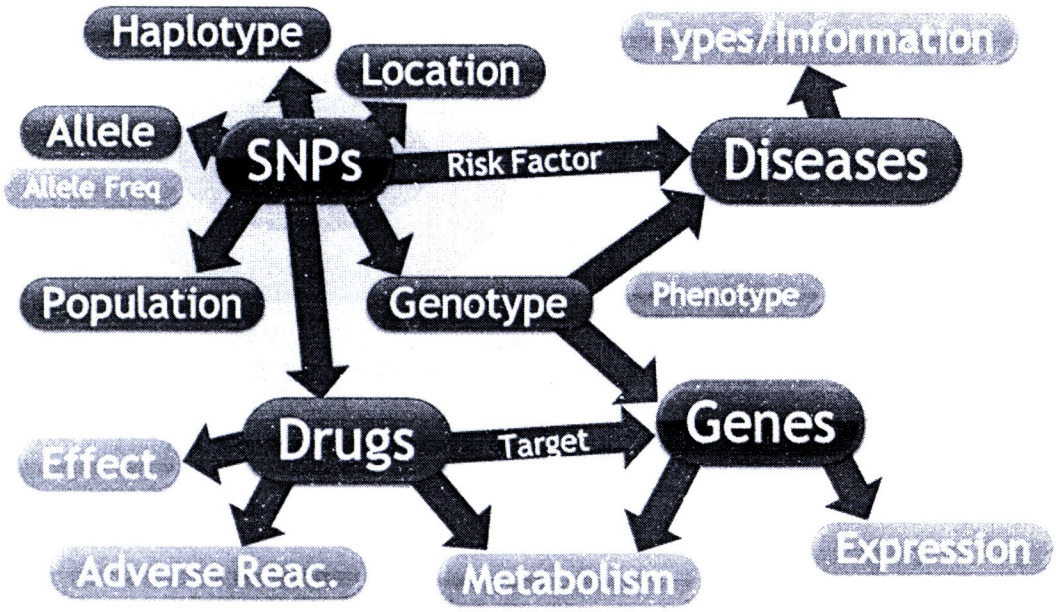



**Figure 4.1** Data linking of SNPs and GWAS study.

## 4.3 Example database and annotation of SNPs and GWAS content pack

An example study for the constructed SNPs and GWAS content pack was the genome wide association study of the high blood pressure traits (Wang, *et al.*, 2009). The general annotation about this study, such as the publication information, the sample size, the replication sample size and the draft information about the associate SNPs and traits were illustrated in the table 4.1. The disease and traits detail information and the SNPs detail information was showed in the table 4.2 and 4.5, respectively. In addition, the experimental and sample details were showed in the table 4.3 and 4.4, respectively. The full SNPs and GWAS content pack are reported in the appendix A.

**Table 4.1** The general information for SNPs and GWAS content pack annotation.

| | |
|---|---|
| **PubMedID** | 19114657 |
| **First Author** | Wang |
| **Publication Date** | 12/29/2008 |
| **Journal** | Proc Natl Acad Sci USA |
| **Link to PubMed** | http://www.ncbi.nlm.nih.gov/pubmed/19114657?ordinalpos=13&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum |
| **Study** | Whole-genome association study identifies STK39 as a hypertension susceptibility gene |

| | |
|---|---|
| **Disease/Trait** | Blood pressure |
| **Initial Sample Size** | 542 individuals |
| **Replication Sample Size** | 6,583 individuals |
| **Region** | 2q24.3 |
| **Reported Gene(s)** | STK39 |
| **Strongest SNP-Risk Allele** | rs6749447-G |
| **SNPs** | rs6749447 |
| **Risk Allele Frequency** | 0.28 |
| **p-Value** | 2.00E-07 |
| **p-Value (text)** | |
| **OR or beta** | 1.9 |
| **95% CI (text)** | [1.2-2.6] mm Hg increase in DBP |
| **Platform [SNPs passing QC]** | Affymetrix [79,447] |
| **CNV** | N |

**Table 4.2** The example diseases or traits annotation.

| | |
|---|---|
| **Disease/Trait** | high blood pressure (hypertension) |
| **Disease definition** | SBP > 140 mmHg (or normal in case of using hypertension medication) |
| | DBP > 90 mmHg (or normal in case of using hypertension medication) |
| **Disease stage** | N/A |

**Table 4.3** The example experimental annotation.

| | |
|---|---|
| **Study** | Genome Wide Association (on autosome) |
| **type of study** | Case / Control and Cohort |
| **Genotype quality control criteria** | Genotype call rate >50% |
| | MAF > 5% |
| | HWE (p-value < 0.001) |
| **SNPs passing QC** | 79,447 |
| **Platform** | Affymetrix 100K |
| **Genotyping Methods** | NR |
| **Statistical Analysis** | Measured genotype approach |

**Table 4.4** The example sample annotation.

| | |
|---|---|
| **Initial Sample Size** | 542 |
| **Initial Sample detail** | Amish Family Diabetes Study (AFDS) |
| **Quality Control** | |
| **Replication Sample Size** | 6583 |
| **Replication Sample detail** | Independent amish and 4 non-amish caucasian samples (Diabetes enetics Initiative, Framingham Heart Study, GenNet, and Hutterites) |
| **Population/Sample details** | |
| **Initial Sample** | Number of sample (use in this study) |

| | |
|---|---|
| **• AFDS** | 542 |
| **Replication Sample** | |
| **• AFDS** | 557 |
| **• HAPI hart study** | 790 |
| **• FHS** | 1345 |
| **• DGI** | 3082 |
| **• Huttrites** | 575 |
| **• GenNet** | 802 |

**Table 4.5** The example SNPs annotation.

| | |
|---|---|
| **SNPs** | rs6749447 |
| **Reference allele** | A/G |
| **SNPs risk allele** | G |
| **Risk allele frequency** | 0.28 |
| **Position** | 168749632 |
| **Region** | 2q24.3 |
| **Reported Gene** | STK39 |
| **p-value** | 2.00E-07 |
| **OR or beta** | 1.9 |
| **95% CI** | [1.2-2.6] mm Hg increase in DBP |
| **CNV** | No |

## 4.4 Emerging correlation

The correlation study was performed by using the constructed SNPs and GWAS content pack and some annotated articles in the SNP and GWAS content pack. The correlation mapping was done by using Cytoscape 2.6.3 (Shannon, *et al.*, 2003). Linkages between genotypes and phenotypes across the study were found. An example correlation between the SNPs and colorectal and prostate cancer traits are shown in the figure 4.2 and the HDL cholesterol, triglycerides, LDL cholesterol and C-reactive protein are shown in the figure 4.3. In addition, the traits-loci correlation is shown in the figure 4.4.

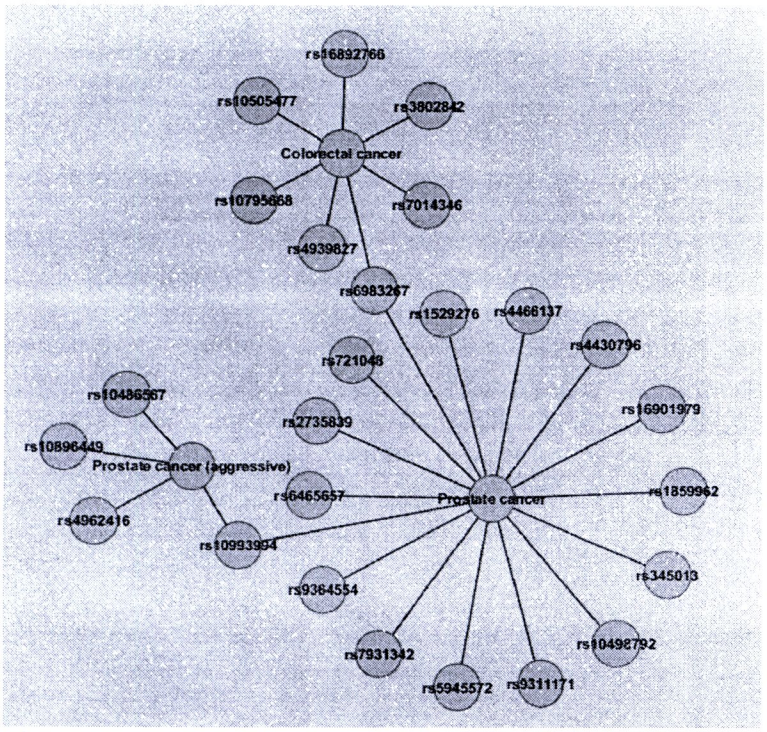



**Figure 4.2** the traits-SNPs correlation between colorectal cancer and prostate cancer was found.

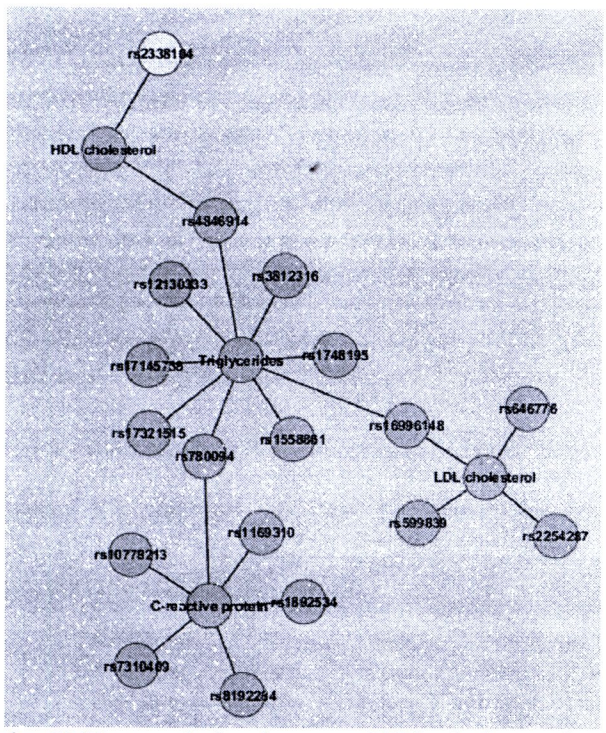



**Figure 4.3** the traits-SNPs correlation between HDL cholesterol, triglycerides, LDL cholesterol and C-reactive protein was found.

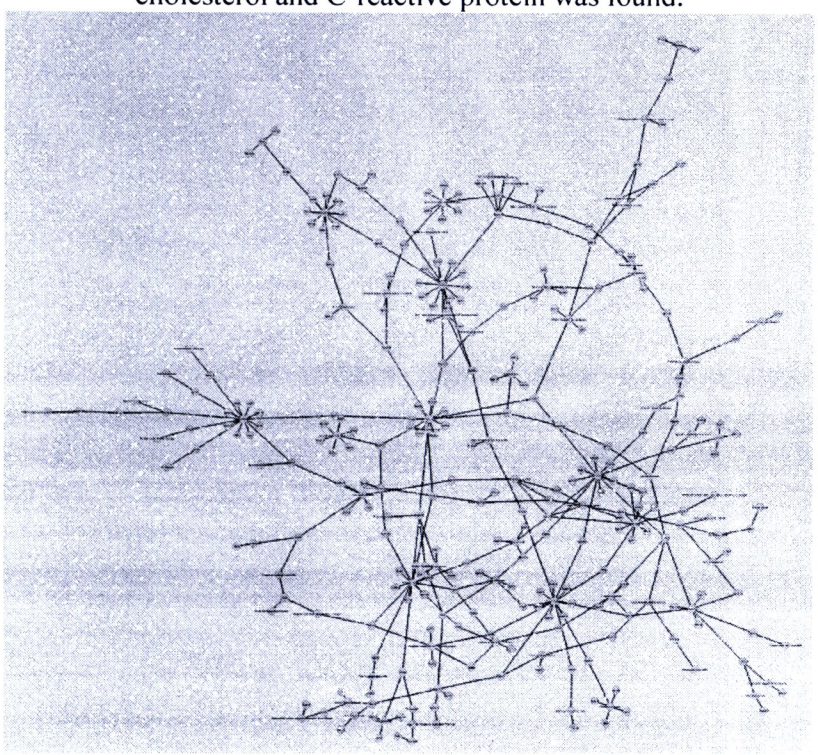


**Figure 4.4** the correlation between traits and the SNPs loci on the chromosome.

## 4.5 Discussions

When the SNPs and GWAS content pack was constructed, the SNPs and GWAS information are standardized into the same format. For example, the information from the genome wide association study are divided in to the general information, trait or disease information, experimental information, sample information and SNPs information. For all annotated information, the same criteria for annotated every fields was used. When the SNPs and GWAS are generalized collected, the constructed SNPs and GWAS content pack can be retrieved genome wide association information easily. Moreover, there are various levels of the emerging information after from the SNPs and GWAS content pack, such as the share genotype (same SNPs) for various disease, the share phenotype (same traits) from various SNPs.

The SNPs and GWAS content pack have been successfully constructed. The constructed SNPs and GWAS content pack are the idea of technology transfer that is the commercialized research knowledge collection. However, the critical comments and suggestions from the researcher are needed in order to improve the structure, the annotated information and the additional analysis.

# CHAPTER 5 CONCLUSION AND RECOMMENDATIONS (II)

## 5.1 Conclusion

The data linkage of SNPs and genome wide association study was constructed in order to find the shared and addition component of SNPs and GWAS content pack. The SNPs and GWAS content pack was successfully constructed from the shared and additional component. Same criteria to collect the genome wide association study into the SNPs and GWAS contest pack were used. Some genome wide association studies were annotated into the SNPs and GWAS content pack. Finally, the correlation studies of SNPs and GWAS content pack were performed to find the emerging linage between traits, loci or SNPs.

## 5.2 Recommendations

This SNPs and GWAS content pack was the primary databases of the SNPs and GWAS that expected to standardized into the same format and easy to visualize. However, the SNPs and GWAS content pack did not include all of the genome wide association study. For improve and generalize the SNPs and GWAS content pack, there are more literature need to include in this content pack.

# REFERENCES (II)


Altshuler, D., Daly, M.J. and Lander, E.S., 2008, "Genetic Mapping in Human Disease", **Science**, Vol. 322, No. 5903, pp. 881-8.

Amos, C.I., Wu, X., Broderick, P., Gorlov, I.P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y.Y., Chen, W.V., Shete, S., Spitz, M.R. and Houlston, R.S., 2008, "Genome-Wide Association Scan of Tag Snps Identifies a Susceptibility Locus for Lung Cancer at 15q25.1", **Nature Genetics**, Vol. 40, No. 5, pp. 616-22.

Anon., 2007, "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls", **Nature**, Vol. 447, No. 7145, pp. 661-78.

Broderick, P., Carvajal-Carmona, L., Pittman, A.M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Spain, S., Sullivan, K., Fielding, S., Jaeger, E., Vijayakrishnan, J., Kemp, Z., Gorman, M., Chandler, I., Papaemmanuil, E., Penegar, S., Wood, W., Sellick, G., Qureshi, M., Teixeira, A., Domingo, E., Barclay, E., Martin, L., Sieber, O., Kerr, D., Gray, R., Peto, J., Cazier, J.B., Tomlinson, I. and Houlston, R.S., 2007, "A Genome-Wide Association Study Shows That Common Alleles of Smad7 Influence Colorectal Cancer Risk", **Nature Genetics**, Vol. 39, No. 11, pp. 1315-7.

Bodmer, W. and Bonilla, C., 2008, "Common and Rare Variants in Multifactorial Susceptibility to Common Diseases", **Nature Genetics**, Vol. 40, No. 6, pp. 695-701.

Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C., Warram, J.H. and Todd, J.A., 2008, "Meta-Analysis of Genome-Wide Association Study Data Identifies Additional Type 1 Diabetes Risk Loci", **Nature Genetics**, Vol. 40, No. 12, pp. 1399-401.

Cupples, L.A., Arruda, H.T., Benjamin, E.J., D'Agostino, R.B., Sr., Demissie, S., DeStefano, A.L., Dupuis, J., Falls, K.M., Fox, C.S., Gottlieb, D.J., Govindaraju, D.R., Guo, C.Y., Heard-Costa, N.L., Hwang, S.J., Kathiresan, S., Kiel, D.P., Laramie, J.M., Larson, M.G., Levy, D., Liu, C.Y., Lunetta, K.L., Mailman, M.D., Manning, A.K., Meigs, J.B., Murabito, J.M., Newton-Cheh, C., O'Connor, G.T., O'Donnell, C.J., Pandey, M., Seshadri, S., Vasan, R.S., Wang, Z.Y., Wilk, J.B., Wolf, P.A., Yang, Q. and Atwood, L.D., 2007, "The Framingham Heart Study 100k Snp Genome-Wide Association Study Resource: Overview of 17 Phenotype Working Group Reports", **BMC medical genetics**, Vol. 8 Suppl 1, No., p. S1.

Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C.S., Bowman, R., Meyer, K.B., Haiman, C.A., Kolonel, L.K., Henderson, B.E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C.Y., Wu, P.E., Wang, H.C., Eccles, D., Evans, D.G., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M.R., Rahman, N., Chenevix-Trench, G., Bojesen, S.E., Nordestgaard, B.G., Axelsson, C.K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K.Y., Noh, D.Y., Ahn, S.H., Hunter, D.J., Hankinson, S.E., Cox, D.G., Hall, P.,

Wedren, S., Liu, J., Low, Y.L., Bogdanova, N., Schurmann, P., Dork, T., Tollenaar, R.A., Jacobi, C.E., Devilee, P., Klijn, J.G., Sigurdson, A.J., Doody, M.M., Alexander, B.H., Zhang, J., Cox, A., Brock, I.W., MacPherson, G., Reed, M.W., Couch, F.J., Goode, E.L., Olson, J.E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R.L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J.L., McCredie, M., Southey, M., Giles, G.G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y.D., Spurdle, A.B., Beesley, J., Chen, X., Mannermaa, A., Kosma, V.M., Kataja, V., Hartikainen, J., Day, N.E., Cox, D.R. and Ponder, B.A., 2007, "Genome-Wide Association Study Identifies Novel Breast Cancer Susceptibility Loci", **Nature**, Vol. 447, No. 7148, pp. 1087-93.

Eeles, R.A., Kote-Jarai, Z., Giles, G.G., Olama, A.A., Guy, M., Jugurnauth, S.K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M., Morrison, J., Field, H.I., Southey, M.C., Severi, G., Donovan, J.L., Hamdy, F.C., Dearnaley, D.P., Muir, K.R., Smith, C., Bagnato, M., Ardern-Jones, A.T., Hall, A.L., O'Brien, L.T., Gehr-Swain, B.N., Wilkinson, R.A., Cox, A., Lewis, S., Brown, P.M., Jhavar, S.G., Tymrakiewicz, M., Lophatananon, A., Bryant, S.L., Horwich, A., Huddart, R.A., Khoo, V.S., Parker, C.C., Woodhouse, C.J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C.S., English, D.R., Hopper, J.L., Neal, D.E. and Easton, D.F., 2008, "Multiple Newly Identified Loci Associated with Prostate Cancer Susceptibility", **Nature Genetics**, Vol. 40, No. 3, pp. 316-21.

Eichler, E.E., Nickerson, D.A., Altshuler, D., Bowcock, A.M., Brooks, L.D., Carter, N.P., Church, D.M., Felsenfeld, A., Guyer, M., Lee, C., Lupski, J.R., Mullikin, J.C., Pritchard, J.K., Sebat, J., Sherry, S.T., Smith, D., Valle, D. and Waterston, R.H., 2007, "Completing the Map of Human Genetic Variation", **Nature**, Vol. 447, No. 7141, pp. 161-5.

Fox, C.S., Heard-Costa, N., Cupples, L.A., Dupuis, J., Vasan, R.S. and Atwood, L.D., 2007, "Genome-Wide Association to Body Mass Index and Waist Circumference: The Framingham Heart Study 100k Project", **BMC medical genetics**, Vol. 8 Suppl 1, No., p. S18.

Frayling, T.M., 2007, "Genome-Wide Association Studies Provide New Insights into Type 2 Diabetes Aetiology", **Nature Reviews. Genetics**, Vol. 8, No. 9, pp. 657-62.

Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W., Shields, B., Harries, L.W., Barrett, J.C., Ellard, S., Groves, C.J., Knight, B., Patch, A.M., Ness, A.R., Ebrahim, S., Lawlor, D.A., Ring, S.M., Ben-Shlomo, Y., Jarvelin, M.R., Sovio, U., Bennett, A.J., Melzer, D., Ferrucci, L., Loos, R.J., Barroso, I., Wareham, N.J., Karpe, F., Owen, K.R., Cardon, L.R., Walker, M., Hitman, G.A., Palmer, C.N., Doney, A.S., Morris, A.D., Smith, G.D., Hattersley, A.T. and McCarthy, M.I., 2007, "A Common Variant in the Fto Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity", **Science**, Vol. 316, No. 5826, pp. 889-94.

Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J., 2009, "Human Genetic Variation and Its Contribution to Complex Traits", **Nature Reviews. Genetics**, Vol. 10, No. 4, pp. 241-51.

Gold, B., Kirchhoff, T., Stefanov, S., Lautenberger, J., Viale, A., Garber, J., Friedman, E., Narod, S., Olshen, A.B., Gregersen, P., Kosarin, K., Olsh, A., Bergeron, J., Ellis, N.A., Klein, R.J., Clark, A.G., Norton, L., Dean, M., Boyd, J. and Offit, K., 2008, "Genome-Wide Association Study Provides Evidence for a Breast Cancer Risk Locus at 6q22.33", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 105, No. 11, pp. 4340-5.

Grant, S.F., Qu, H.Q., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Taback, S.P., Frackelton, E.C., Eckert, A.W., Annaiah, K., Lawson, M.L., Otieno, F.G., Santa, E., Shaner, J.L., Smith, R.M., Skraban, R., Imielinski, M., Chiavacci, R.M., Grundmeier, R.W., Stanley, C.A., Kirsch, S.E., Waggott, D., Paterson, A.D., Monos, D.S., Polychronakos, C. and Hakonarson, H., 2009, "Follow-up Analysis of Genome-Wide Association Data Identifies Novel Loci for Type 1 Diabetes", **Diabetes**, Vol. 58, No. 1, pp. 290-5.

Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L.T., Gudbjartsson, D., Helgason, A., Rafnar, T., Bergthorsson, J.T., Agnarsson, B.A., Baker, A., Sigurdsson, A., Benediktsdottir, K.R., Jakobsdottir, M., Xu, J., Blondal, T., Kostic, J., Sun, J., Ghosh, S., Stacey, S.N., Mouy, M., Saemundsdottir, J., Backman, V.M., Kristjansson, K., Tres, A., Partin, A.W., Albers-Akkers, M.T., Godino-Ivan Marcos, J., Walsh, P.C., Swinkels, D.W., Navarrete, S., Isaacs, S.D., Aben, K.K., Graif, T., Cashy, J., Ruiz-Echarri, M., Wiley, K.E., Suarez, B.K., Witjes, J.A., Frigge, M., Ober, C., Jonsson, E., Einarsson, G.V., Mayordomo, J.I., Kiemeney, L.A., Isaacs, W.B., Catalona, W.J., Barkardottir, R.B., Gulcher, J.R., Thorsteinsdottir, U., Kong, A. and Stefansson, K., 2007a, "Genome-Wide Association Study Identifies a Second Prostate Cancer Susceptibility Variant at 8q24", **Nature Genetics**, Vol. 39, No. 5, pp. 631-7.

Gudmundsson, J., Sulem, P., Rafnar, T., Bergthorsson, J.T., Manolescu, A., Gudbjartsson, D., Agnarsson, B.A., Sigurdsson, A., Benediktsdottir, K.R., Blondal, T., Jakobsdottir, M., Stacey, S.N., Kostic, J., Kristinsson, K.T., Birgisdottir, B., Ghosh, S., Magnusdottir, D.N., Thorlacius, S., Thorleifsson, G., Zheng, S.L., Sun, J., Chang, B.L., Elmore, J.B., Breyer, J.P., McReynolds, K.M., Bradley, K.M., Yaspan, B.L., Wiklund, F., Stattin, P., Lindstrom, S., Adami, H.O., McDonnell, S.K., Schaid, D.J., Cunningham, J.M., Wang, L., Cerhan, J.R., St Sauver, J.L., Isaacs, S.D., Wiley, K.E., Partin, A.W., Walsh, P.C., Polo, S., Ruiz-Echarri, M., Navarrete, S., Fuertes, F., Saez, B., Godino, J., Weijerman, P.C., Swinkels, D.W., Aben, K.K., Witjes, J.A., Suarez, B.K., Helfand, B.T., Frigge, M.L., Kristjansson, K., Ober, C., Jonsson, E., Einarsson, G.V., Xu, J., Gronberg, H., Smith, J.R., Thibodeau, S.N., Isaacs, W.B., Catalona, W.J., Mayordomo, J.I., Kiemeney, L.A., Barkardottir, R.B., Gulcher, J.R., Thorsteinsdottir, U., Kong, A. and Stefansson, K., 2008, "Common Sequence Variants on 2p15 and Xp11.22 Confer Susceptibility to Prostate Cancer", **Nature Genetics**, Vol. 40, No. 3, pp. 281-3.

Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J.T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B.A., Baker, A., Sigurdsson, A., Benediktsdottir, K.R., Jakobsdottir, M., Blondal, T., Stacey, S.N., Helgason, A., Gunnarsdottir, S., Olafsdottir, A., Kristinsson, K.T., Birgisdottir, B., Ghosh, S., Thorlacius, S., Magnusdottir, D., Stefansdottir, G., Kristjansson, K., Bagger, Y., Wilensky, R.L., Reilly, M.P., Morris, A.D., Kimber, C.H., Adeyemo, A., Chen, Y., Zhou, J., So, W.Y., Tong, P.C., Ng, M.C., Hansen, T., Andersen, G., Borch-Johnsen,

K., Jorgensen, T., Tres, A., Fuertes, F., Ruiz-Echarri, M., Asin, L., Saez, B., van Boven, E., Klaver, S., Swinkels, D.W., Aben, K.K., Graif, T., Cashy, J., Suarez, B.K., van Vierssen Trip, O., Frigge, M.L., Ober, C., Hofker, M.H., Wijmenga, C., Christiansen, C., Rader, D.J., Palmer, C.N., Rotimi, C., Chan, J.C., Pedersen, O., Sigurdsson, G., Benediktsson, R., Jonsson, E., Einarsson, G.V., Mayordomo, J.I., Catalona, W.J., Kiemeney, L.A., Barkardottir, R.B., Gulcher, J.R., Thorsteinsdottir, U., Kong, A. and Stefansson, K., 2007b, "Two Variants on Chromosome 17 Confer Prostate Cancer Risk, and the One in Tcf2 Protects against Type 2 Diabetes", **Nature Genetics**, Vol. 39, No. 8, pp. 977-83.

Hakonarson, H., Grant, S.F., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Casalunovo, T., Taback, S.P., Frackelton, E.C., Lawson, M.L., Robinson, L.J., Skraban, R., Lu, Y., Chiavacci, R.M., Stanley, C.A., Kirsch, S.E., Rappaport, E.F., Orange, J.S., Monos, D.S., Devoto, M., Qu, H.Q. and Polychronakos, C., 2007, "A Genome-Wide Association Study Identifies Kiaa0350 as a Type 1 Diabetes Gene", **Nature**, Vol. 448, No. 7153, pp. 591-4.

Hakonarson, H., Qu, H.Q., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Casalunovo, T., Taback, S.P., Frackelton, E.C., Eckert, A.W., Annaiah, K., Lawson, M.L., Otieno, F.G., Santa, E., Shaner, J.L., Smith, R.M., Onyiah, C.C., Skraban, R., Chiavacci, R.M., Robinson, L.J., Stanley, C.A., Kirsch, S.E., Devoto, M., Monos, D.S., Grant, S.F. and Polychronakos, C., 2008, "A Novel Susceptibility Locus for Type 1 Diabetes on Chr12q13 Identified by a Genome-Wide Association Study", **Diabetes**, Vol. 57, No. 4, pp. 1143-6.

Hardy, J. and Singleton, A., 2009, "Genomewide Association Studies and Human Disease", **New England Journal of Medicine**, Vol. 360, No. 17, pp. 1759-68.

Hung, R.J., McKay, J.D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., Chen, C., Goodman, G., Field, J.K., Liloglou, T., Xinarianos, G., Cassidy, A., McLaughlin, J., Liu, G., Narod, S., Krokan, H.E., Skorpen, F., Elvestad, M.B., Hveem, K., Vatten, L., Linseisen, J., Clavel-Chapelon, F., Vineis, P., Bueno-de-Mesquita, H.B., Lund, E., Martinez, C., Bingham, S., Rasmuson, T., Hainaut, P., Riboli, E., Ahrens, W., Benhamou, S., Lagiou, P., Trichopoulos, D., Holcatova, I., Merletti, F., Kjaerheim, K., Agudo, A., Macfarlane, G., Talamini, R., Simonato, L., Lowry, R., Conway, D.I., Znaor, A., Healy, C., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M. and Brennan, P., 2008, "A Susceptibility Locus for Lung Cancer Maps to Nicotinic Acetylcholine Receptor Subunit Genes on 15q25", **Nature**, Vol. 452, No. 7187, pp. 633-7.

Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W.C., Colditz, G.A., Ziegler, R.G., Berg, C.D., Buys, S.S., McCarty, C.A., Feigelson, H.S., Calle, E.E., Thun, M.J., Hayes, R.B., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Jr., Hoover, R.N., Thomas, G. and Chanock, S.J., 2007, "A Genome-Wide Association Study Identifies Alleles in Fgfr2 Associated with Risk of Sporadic Postmenopausal Breast Cancer", **Nature Genetics**, Vol. 39, No. 7, pp. 870-4.

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N.P., Rieder, M.J., Cooper, G.M., Roos, C., Voight, B.F., Havulinna, A.S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E.S., Ordovas, J.M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M.R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D.M. and Orho-Melander, M., 2008, "Six New Loci Associated with Blood Low-Density Lipoprotein Cholesterol, High-Density Lipoprotein Cholesterol or Triglycerides in Humans", **Nature Genetics**, Vol. 40, No. 2, pp. 189-97.

Kibriya, M.G., Jasmine, F., Argos, M., Andrulis, I.L., John, E.M., Chang-Claude, J. and Ahsan, H., 2009, "A Pilot Genome-Wide Association Study of Early-Onset Breast Cancer", **Breast Cancer Research and Treatment**, Vol. 114, No. 3, pp. 463-77.

Kingsmore, S.F., Lindquist, I.E., Mudge, J., Gessler, D.D. and Beavis, W.D., 2008, "Genome-Wide Association Studies: Progress and Potential for Drug Discovery and Development", **Nature Reviews. Drug Discovery**, Vol. 7, No. 3, pp. 221-30.

Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gomez Perez, F.J., Frazer, K.A., Elliott, P., Scott, J., Milos, P.M., Cox, D.R. and Thompson, J.F., 2008, "Genome-Wide Scan Identifies Variation in Mlxipl Associated with Plasma Triglycerides", **Nature Genetics**, Vol. 40, No. 2, pp. 149-51.

Levy, D., Larson, M.G., Benjamin, E.J., Newton-Cheh, C., Wang, T.J., Hwang, S.J., Vasan, R.S. and Mitchell, G.F., 2007, "Framingham Heart Study 100k Project: Genome-Wide Associations for Blood Pressure and Arterial Stiffness", **BMC medical genetics**, Vol. 8 Suppl 1, No., p. S3.

Loos, R.J., Lindgren, C.M., Li, S., Wheeler, E., Zhao, J.H., Prokopenko, I., Inouye, M., Freathy, R.M., Attwood, A.P., Beckmann, J.S., Berndt, S.I., Jacobs, K.B., Chanock, S.J., Hayes, R.B., Bergmann, S., Bennett, A.J., Bingham, S.A., Bochud, M., Brown, M., Cauchi, S., Connell, J.M., Cooper, C., Smith, G.D., Day, I., Dina, C., De, S., Dermitzakis, E.T., Doney, A.S., Elliott, K.S., Elliott, P., Evans, D.M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C.J., Gwilliam, R., Hadley, D., Hall, A.S., Hattersley, A.T., Hebebrand, J., Heid, I.M., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H.E., Herrera, B., Hinney, A., Hunt, S.E., Jarvelin, M.R., Johnson, T., Jolley, J.D., Karpe, F., Keniry, A., Khaw, K.T., Luben, R.N., Mangino, M., Marchini, J., McArdle, W.L., McGinnis, R., Meyre, D., Munroe, P.B., Morris, A.D., Ness, A.R., Neville, M.J., Nica, A.C., Ong, K.K., O'Rahilly, S., Owen, K.R., Palmer, C.N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J.C., Rayner, N.W., Ring, S.M., Sandhu, M.S., Scherag, A., Sims, M.A., Song, K., Soranzo, N., Speliotes, E.K., Syddall, H.E., Teichmann, S.A., Timpson, N.J., Tobias, J.H., Uda, M., Vogel, C.I., Wallace, C., Waterworth, D.M., Weedon, M.N., Willer, C.J., Wraight, Yuan, X., Zeggini, E., Hirschhorn, J.N., Strachan, D.P., Ouwehand, W.H., Caulfield, M.J., Samani, N.J., Frayling, T.M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M.I., Wareham, N.J., Barroso, I., Kraft, P., Hankinson, S.E., Hunter, D.J., Hu, F.B., Lyon, H.N., Voight, B.F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G.R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A.U., Tuomilehto, J., Collins, F.S., Boehnke, M. and Mohlke, K.L., 2008, "Common Variants near Mc4r Are Associated with Fat Mass, Weight and Risk of Obesity", **Nature Genetics**, Vol. 40, No. 6, pp. 768-75.

Manolio, T.A., Bailey-Wilson, J.E. and Collins, F.S., 2006, "Genes, Environment and the Value of Prospective Cohort Studies", **Nature Reviews. Genetics**, Vol. 7, No. 10, pp. 812-20.

Murabito, J.M., Rosenberg, C.L., Finger, D., Kreger, B.E., Levy, D., Splansky, G.L., Antman, K. and Hwang, S.J., 2007, "A Genome-Wide Association Study of Breast and Prostate Cancer in the Nhlbi's Framingham Heart Study", **BMC medical genetics**, Vol. 8 Suppl 1, No., p. S6.

Pearson, T.A. and Manolio, T.A., 2008, "How to Interpret a Genome-Wide Association Study", **JAMA : the Journal of the American Medical Association**, Vol. 299, No. 11, pp. 1335-44.

Reiner, A.P., Barber, M.J., Guan, Y., Ridker, P.M., Lange, L.A., Chasman, D.I., Walston, J.D., Cooper, G.M., Jenny, N.S., Rieder, M.J., Durda, J.P., Smith, J.D., Novembre, J., Tracy, R.P., Rotter, J.I., Stephens, M., Nickerson, D.A. and Krauss, R.M., 2008, "Polymorphisms of the Hnf1a Gene Encoding Hepatocyte Nuclear Factor-1 Alpha Are Associated with C-Reactive Protein", **American Journal of Human Genetics**, Vol. 82, No. 5, pp. 1193-201.

Ridker, P.M., Chasman, D.I., Zee, R.Y., Parker, A., Rose, L., Cook, N.R. and Buring, J.E., 2008, "Rationale, Design, and Methodology of the Women's Genome Health Study: A Genome-Wide Association Study of More Than 25,000 Initially Healthy American Women", **Clinical Chemistry**, Vol. 54, No. 2, pp. 249-55.

Ridker, P.M., Pare, G., Parker, A., Zee, R.Y., Danik, J.S., Buring, J.E., Kwiatkowski, D., Cook, N.R., Miletich, J.P. and Chasman, D.I., 2008, "Loci Related to Metabolic-Syndrome Pathways Including Lepr,Hnf1a, Il6r, and Gckr Associate with Plasma C-Reactive Protein: The Women's Genome Health Study", **American Journal of Human Genetics**, Vol. 82, No. 5, pp. 1185-92.

Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., Song, K., Yuan, X., Johnson, T., Ashford, S., Inouye, M., Luben, R., Sims, M., Hadley, D., McArdle, W., Barter, P., Kesaniemi, Y.A., Mahley, R.W., McPherson, R., Grundy, S.M., Bingham, S.A., Khaw, K.T., Loos, R.J., Waeber, G., Barroso, I., Strachan, D.P., Deloukas, P., Vollenweider, P., Wareham, N.J. and Mooser, V., 2008, "Ldl-Cholesterol Concentrations: A Genome-Wide Association Study", **Lancet**, Vol. 371, No. 9611, pp. 483-91.

Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J., Hughes, T.E., Groop, L., Altshuler, D., Almgren, P., Florez, J.C., Meyer, J., Ardlie, K., Bengtsson Bostrom, K., Isomaa, B., Lettre, G., Lindblad, U., Lyon, H.N., Melander, O., Newton-Cheh, C., Nilsson, P., Orho-Melander, M., Rastam, L., Speliotes, E.K., Taskinen, M.R., Tuomi, T., Guiducci, C., Berglund, A., Carlson, J., Gianniny, L., Hackett, R., Hall, L., Holmkvist, J., Laurila, E., Sjogren, M., Sterner, M., Surti, A., Svensson, M., Tewhey, R., Blumenstiel, B., Parkin, M., Defelice, M., Barry, R., Brodeur, W., Camarata, J., Chia, N., Fava, M., Gibbons, J., Handsaker, B., Healy, C., Nguyen, K., Gates, C., Sougnez, C., Gage, D., Nizzari, M., Gabriel, S.B., Chirn, G.W., Ma, Q., Parikh, H., Richardson, D., Ricke, D.

and Purcell, S., 2007, "Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels", **Science**, Vol. 316, No. 5829, pp. 1331-6.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S. and Boehnke, M., 2007, "A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants", **Science**, Vol. 316, No. 5829, pp. 1341-5.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., 2003, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", **Genome Research**, Vol. 13, No. 11, pp. 2498-504.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. and Froguel, P., 2007, "A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes", **Nature**, Vol. 445, No. 7130, pp. 881-5.

Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., Aben, K.K., Strobbe, L.J., Albers-Akkers, M.T., Swinkels, D.W., Henderson, B.E., Kolonel, L.N., Le Marchand, L., Millastre, E., Andres, R., Godino, J., Garcia-Prats, M.D., Polo, E., Tres, A., Mouy, M., Saemundsdottir, J., Backman, V.M., Gudmundsson, L., Kristjansson, K., Bergthorsson, J.T., Kostic, J., Frigge, M.L., Geller, F., Gudbjartsson, D., Sigurdsson, H., Jonsdottir, T., Hrafnkelsson, J., Johannsson, J., Sveinsson, T., Myrdal, G., Grimsson, H.N., Jonsson, T., von Holst, S., Werelius, B., Margolin, S., Lindblom, A., Mayordomo, J.I., Haiman, C.A., Kiemeney, L.A., Johannsson, O.T., Gulcher, J.R., Thorsteinsdottir, U., Kong, A. and Stefansson, K., 2007, "Common Variants on Chromosomes 2q35 and 16q12 Confer Susceptibility to Estrogen Receptor-Positive Breast Cancer", **Nature Genetics**, Vol. 39, No. 7, pp. 865-9.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., Bjarnason, H., Ng, M.C., Hansen, T., Bagger, Y., Wilensky, R.L., Reilly, M.P., Adeyemo, A., Chen, Y., Zhou, J., Gudnason, V., Chen, G., Huang, H., Lashley, K., Doumatey, A., So, W.Y., Ma, R.C., Andersen, G., Borch-Johnsen, K., Jorgensen, T., van Vliet-Ostaptchouk, J.V., Hofker, M.H., Wijmenga, C., Christiansen, C., Rader, D.J., Rotimi, C., Gurney, M., Chan, J.C., Pedersen, O., Sigurdsson, G., Gulcher, J.R., Thorsteinsdottir, U., Kong, A. and Stefansson, K., 2007, "A Variant in Cdkal1 Influences Insulin Response and Risk of Type 2 Diabetes", **Nature Genetics**, Vol. 39, No. 6, pp. 770-5.

Tenesa, A., Farrington, S.M., Prendergast, J.G., Porteous, M.E., Walker, M., Haq, N., Barnetson, R.A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., Semple, C., Clark, A.J., Reid, F.J., Smith, L.A., Kavoussanakis, K., Koessler, T., Pharoah, P.D., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Volzke, H., Schmidt, C.O., Hampe, J., Chang-Claude, J., Hoffmeister, M., Brenner, H., Wilkening, S., Canzian, F., Capella, G., Moreno, V., Deary, I.J., Starr, J.M., Tomlinson, I.P., Kemp, Z., Howarth, K., Carvajal-Carmona, L., Webb, E., Broderick, P., Vijayakrishnan, J., Houlston, R.S., Rennert, G., Ballinger, D., Rozek, L., Gruber, S.B., Matsuda, K., Kidokoro, T., Nakamura, Y., Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Montpetit, A., Hudson, T.J., Gallinger, S., Campbell, H. and Dunlop, M.G., 2008, "Genome-Wide Association Scan Identifies a Colorectal Cancer Susceptibility Locus on 11q23 and Replicates Risk Loci at 8q24 and 18q21", **Nature Genetics**, Vol. 40, No. 5, pp. 631-7.

Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., Lowe, C.E., Szeszko, J.S., Hafler, J.P., Zeitels, L., Yang, J.H., Vella, A., Nutland, S., Stevens, H.E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L.J., Healy, B., Burren, O.S., Lam, A.A., Ovington, N.R., Allen, J., Adlem, E., Leung, H.T., Wallace, C., Howson, J.M., Guja, C., Ionescu-Tirgoviste, C., Simmonds, M.J., Heward, J.M., Gough, S.C., Dunger, D.B., Wicker, L.S. and Clayton, D.G., 2007, "Robust Associations of Four New Chromosome Regions from Genome-Wide Analyses of Type 1 Diabetes", **Nature Genetics**, Vol. 39, No. 7, pp. 857-64.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., Howarth, K., Silver, A., Atkin, W., Muir, K., Logan, R., Kerr, D., Johnstone, E., Sieber, O., Gray, R., Thomas, H., Peto, J., Cazier, J.B. and Houlston, R., 2007, "A Genome-Wide Association Scan of Tag Snps Identifies a Susceptibility Variant for Colorectal Cancer at 8q24.21", **Nature Genetics**, Vol. 39, No. 8, pp. 984-8.

Tomlinson, I.P., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A.M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., Jaeger, E., Fielding, S., Rowan, A., Vijayakrishnan, J., Domingo, E., Chandler, I., Kemp, Z., Qureshi, M., Farrington, S.M., Tenesa, A., Prendergast, J.G., Barnetson, R.A., Penegar, S., Barclay, E., Wood, W., Martin, L., Gorman, M., Thomas, H., Peto, J., Bishop, D.T., Gray, R., Maher, E.R., Lucassen, A., Kerr, D., Evans, D.G., Schafmayer, C., Buch, S., Volzke, H., Hampe, J., Schreiber, S., John, U., Koessler, T., Pharoah, P., van Wezel, T., Morreau, H., Wijnen, J.T., Hopper, J.L., Southey, M.C., Giles, G.G., Severi, G., Castellvi-Bel, S., Ruiz-Ponte, C., Carracedo, A., Castells, A., Forsti, A., Hemminki, K., Vodicka, P., Naccarati, A., Lipton, L., Ho, J.W., Cheng, K.K., Sham, P.C., Luk, J., Agundez, J.A., Ladero, J.M., de la Hoya, M., Caldes, T., Niittymaki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Cazier, J.B., Campbell, H., Dunlop, M.G. and Houlston, R.S., 2008, "A Genome-Wide Association Study Identifies Colorectal Cancer Susceptibility Loci on Chromosomes 10p14 and 8q23.3", **Nature Genetics**, Vol. 40, No. 5, pp. 623-30.

Wallace, C., Newhouse, S.J., Braund, P., Zhang, F., Tobin, M., Falchi, M., Ahmadi, K., Dobson, R.J., Marcano, A.C., Hajat, C., Burton, P., Deloukas, P., Brown, M., Connell, J.M., Dominiczak, A., Lathrop, G.M., Webster, J., Farrall, M., Spector, T., Samani, N.J., Caulfield, M.J. and Munroe, P.B., 2008, "Genome-Wide Association Study

Identifies Genes for Biomarkers of Cardiovascular Disease: Serum Urate and Dyslipidemia", **American Journal of Human Genetics**, Vol. 82, No. 1, pp. 139-49.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W.V., Spitz, M.R., Eisen, T., Amos, C.I. and Houlston, R.S., 2008, "Common 5p15.33 and 6p21.33 Variants Influence Lung Cancer Risk", **Nature Genetics**, Vol. 40, No. 12, pp. 1407-9.

Wang, Y., O'Connell, J.R., McArdle, P.F., Wade, J.B., Dorff, S.E., Shah, S.J., Shi, X., Pan, L., Rampersaud, E., Shen, H., Kim, J.D., Subramanya, A.R., Steinle, N.I., Parsa, A., Ober, C.C., Welling, P.A., Chakravarti, A., Weder, A.B., Cooper, R.S., Mitchell, B.D., Shuldiner, A.R. and Chang, Y.P., 2009, "From the Cover: Whole-Genome Association Study Identifies Stk39 as a Hypertension Susceptibility Gene", **Proceedings of the National Academy of Sciences of the United States of America**, Vol. 106, No. 1, pp. 226-31.

Willer, C.J., Sanna, S., Jackson, A.U., Scuteri, A., Bonnycastle, L.L., Clarke, R., Heath, S.C., Timpson, N.J., Najjar, S.S., Stringham, H.M., Strait, J., Duren, W.L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A.J., Morken, M.A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W.M., Li, Y., Scott, L.J., Scheet, P.A., Sundvall, J., Watanabe, R.M., Nagaraja, R., Ebrahim, S., Lawlor, D.A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A.R., Collins, R., Bergman, R.N., Uda, M., Tuomilehto, J., Cao, A., Collins, F.S., Lakatta, E., Lathrop, G.M., Boehnke, M., Schlessinger, D., Mohlke, K.L. and Abecasis, G.R., 2008, "Newly Identified Loci That Influence Lipid Concentrations and Risk of Coronary Artery Disease", **Nature Genetics**, Vol. 40, No. 2, pp. 161-9.

Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B.J., Calle, E.E., Feigelson, H.S., Thun, M.J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Giovannucci, E., Willett, W.C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G.L., Gelmann, E.P., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Jr., Hoover, R., Hunter, D.J., Chanock, S.J. and Thomas, G., 2007, "Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24", **Nature Genetics**, Vol. 39, No. 5, pp. 645-9.

Zanke, B.W., Greenwood, C.M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S.M., Prendergast, J., Olschwang, S., Chiang, T., Crowdy, E., Ferretti, V., Laflamme, P., Sundararajan, S., Roumy, S., Olivier, J.F., Robidoux, F., Sladek, R., Montpetit, A., Campbell, P., Bezieau, S., O'Shea, A.M., Zogopoulos, G., Cotterchio, M., Newcomb, P., McLaughlin, J., Younghusband, B., Green, R., Green, J., Porteous, M.E., Campbell, H., Blanche, H., Sahbatou, M., Tubacher, E., Bonaiti-Pellie, C., Buecher, B., Riboli, E., Kury, S., Chanock, S.J., Potter, J., Thomas, G., Gallinger, S., Hudson, T.J. and Dunlop, M.G., 2007, "Genome-Wide Association Scan Identifies a Colorectal Cancer Susceptibility Locus on Chromosome 8q24", **Nature Genetics**, Vol. 39, No. 8, pp. 989-94.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J.,

Deodhar, P., Ding, C.J., Doney, A.S., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N., Payne, F., Perry, J.R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J., Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., McCarthy, M.I., Boehnke, M. and Altshuler, D., 2008, "Meta-Analysis of Genome-Wide Association Data and Large-Scale Replication Identifies Additional Susceptibility Loci for Type 2 Diabetes", **Nature Genetics**, Vol. 40, No. 5, pp. 638-45.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M., Barrett, J.C., Shields, B., Morris, A.P., Ellard, S., Groves, C.J., Harries, L.W., Marchini, J.L., Owen, K.R., Knight, B., Cardon, L.R., Walker, M., Hitman, G.A., Morris, A.D., Doney, A.S., McCarthy, M.I. and Hattersley, A.T., 2007, "Replication of Genome-Wide Association Signals in Uk Samples Reveals Risk Loci for Type 2 Diabetes", **Science**, Vol. 316, No. 5829, pp. 1336-41.

# APPENDIX A

Full SNPs and GWAS content pack

**Table A.1** Full SNPs and GWAS content pack.

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Cooper | Type 1 diabetes | 2p23.3 | Intergenic | rs2165738 | 0.27 | 0.000004 | (Cooper, J.D., Smyth, D.J. *et al.*, 2008) |
| Cooper | Type 1 diabetes | 6q15 | BACH2 | rs11755527 | 0.47 | 5.00E-12 | |
| Cooper | Type 1 diabetes | 10p15.1 | PRKCQ | rs947474 | 0.19 | 4.00E-09 | |
| Cooper | Type 1 diabetes | 15q25.1 | CTSH | rs3825932 | 0.32 | 3.00E-15 | |
| Cooper | Type 1 diabetes | 22q13.1 | C1QTNF6 | rs229541 | 0.43 | 2.00E-08 | |
| Wang | Lung cancer | 6p21.33 | BAT3MSH5 | rs3117582 | NR | 5.00E-10 | (Wang, Y., Broderick, P. *et al.*, 2008) |
| Wang | Lung cancer | 5p15.33 | CLPTM1L | rs401681 | NR | 8.00E-09 | |
| Grant | Type 1 diabetes | 6q15 | BACH2 | rs3757247 | NR | 1.00E-06 | (Grant, S.F., Qu, H.Q. *et al.*, 2009) |
| Grant | Type 1 diabetes | 15q14 | RASGRP1 | rs8035957 | NR | 4E-06 | |
| Grant | Type 1 diabetes | 1p22.3 | EDG7 | rs1983853 | NR | 2E-06 | |
| Grant | Type 1 diabetes | 21q22.3 | UBASH3A | rs9976767 | NR | 2.00E-08 | |
| Grant | Type 1 diabetes | 9p24.2 | GLIS3 | rs10758593 | NR | 0.000003 | |
| Kibriya | Breast cancer | 2q37.1 | GLG1 | rs10871290 | 0.34 | 4.00E-07 | (Kibriya, M.G., Jasmine, F. *et al.*, 2009) |
| Loos | Body mass index | 18q21.32 | MC4R | rs17782313 | 0.24 | 3.00E-15 | (Loos, R.J., Lindgren, C.M. *et al.*, 2008) |
| Reiner | C-reactive protein | 12q24.31 | HNF1A | rs1169310 | 0.38 | 2.00E-08 | (Reiner, A.P., Barber, M.J. *et al.*, 2008) |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Ridker | C-reactive protein | 1p31.3 | LEPR | rs1892534 | NR | 7.00E-21 | (Ridker, P.M., Pare, G. *et al.*, 2008) |
| Ridker | C-reactive protein | 12q23.2 | Unknown | rs10778213 | NR | 1.00E-10 | |
| Ridker | C-reactive protein | 1q21.3 | IL6R | rs8192284 | NR | 2.00E-08 | |
| Ridker | C-reactive protein | 12q24.31 | HNF1A | rs7310409 | NR | 7.00E-17 | |
| Ridker | C-reactive protein | 2p23.3 | GCKR | rs780094 | NR | 7.00E-15 | |
| Amos | Lung cancer | 15q25.1 | CHRNA3, CHRNA5, PSMA4, LOC123688 | rs8034191 | NR | 3.00E-18 | (Amos, C.I., Wu, X. *et al.*, 2008) |
| Amos | Lung cancer | 3q28 | IL1RAP | rs7626795 | NR | 8E-06 | |
| Amos | Lung cancer | 1q23.2 | CRP | rs2808630 | NR | 7E-06 | |
| Hung | Lung cancer | 15q25.1 | CHRNA3,CHRNA5, CHRNB4, PSMA4, LOC123688 | rs8034191 | 0.34 | 5.00E-20 | (Hung, R.J., McKay, J.D. *et al.*, 2008) |
| Tenesa | Colorectal cancer | 8q24.21 | POU5FIP1, HsG57825, DQ515897 | rs7014346 | 0.18 | 9.00E-26 | (Tenesa, A., Farrington, S.M. *et al.*, 2008) |
| Tenesa | Colorectal cancer | 18q21.1 | SMAD7 | rs4939827 | 0.17 | 8.00E-28 | |
| Tenesa | Colorectal cancer | 11q23.1 | Intergenic | rs3802842 | 0.43 | 6.00E-10 | |
| Tomlinson | Colorectal cancer | 10p14 | Intergenic | rs10795668 | 0.67 | 3.00E-13 | (Tomlinson, I.P., Webb, E. *et al.*, 2008) |
| Tomlinson | Colorectal cancer | 8q23.3 | EIF3H | rs16892766 | 0.07 | 3.00E-18 | |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Zeggini | Type 2 diabetes | 7p15.1 | JAZF1 | rs864745 | 0.5 | 5.00E-14 | (Zeggini, E., Scott, L.J. *et al.*, 2008) |
| Zeggini | Type 2 diabetes | 2p21 | THADA | rs7578597 | 0.9 | 1.00E-09 | |
| Zeggini | Type 2 diabetes | 3p14.1 | ADAMTS9 | rs4607103 | 0.76 | 1.00E-08 | |
| Zeggini | Type 2 diabetes | 10p13 | CDC123,CAMK1D | rs12779790 | 0.18 | 1.00E-10 | |
| Zeggini | Type 2 diabetes | 12q21.1 | TSPAN8,LGR5 | rs7961581 | 0.27 | 1.00E-09 | |
| Gold | Breast cancer | 6q22.33 | ECHDC1,RNF146 | rs2180341 | 0.21 | 3.00E-08 | (Gold, B., Kirchhoff, T. *et al.*, 2008) |
| Eeles | Prostate cancer | 10q11.23 | MSMB | rs10993994 | 0.4 | 9.00E-29 | (Eeles, R.A., Kote-Jarai, Z. *et al.*, 2008) |
| Eeles | Prostate cancer | 6q25.3 | SLC22A3 | rs9364554 | 0.29 | 6.00E-10 | |
| Eeles | Prostate cancer | 7q21.3 | LMTK2 | rs6465657 | 0.46 | 1.00E-09 | |
| Eeles | Prostate cancer | 11q13.2 | Intergenic | rs7931342 | 0.51 | 2.00E-12 | |
| Eeles | Prostate cancer | 19q13.33 | KLK3 | rs2735839 | 0.85 | 2.00E-18 | |
| Gudmundsson | Prostate cancer | Xp11.22 | NUDT10, NUDT11, LOC340602, GSPT2, MAGED1 | rs5945572 | 0.35 | 4.00E-13 | (Gudmundsson, J., Sulem, P. *et al.*, 2008) |
| Gudmundsson | Prostate cancer | 2p15 | EHBP1 | rs721048 | 0.19 | 8.00E-09 | |
| Sandhu | LDL cholesterol | 1p13.3 | CELSR2,PSRC1 | rs599839 | 0.19 | 1.00E-33 | (Sandhu, M.S., Waterworth, D.M. *et al.*, 2008) |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Hakonarson | Type 1 diabetes | 12q13.2 | RAB5B, SUOX, IKZF4, ERBB3, CDK2 | rs1701704 | 0.35 | 9.00E-10 | (Hakonarson, H., Qu, H.Q. *et al.*, 2008) |
| Kathiresan | HDL cholesterol | 1q42.13 | GALNT2 | rs4846914 | 0.4 | 2.00E-13 | (Kathiresan, S., Melander, O. *et al.*, 2008) |
| Kathiresan | LDL cholesterol | 1p13.3 | CELSR2,PSRC1,SORT1 | rs646776 | 0.24 | 3.00E-29 | |
| Kathiresan | LDL cholesterol | 19p13.11 | CILP2, PBX4 | rs16996148 | 0.9 | 3.00E-08 | |
| Kathiresan | Triglycerides | 7q11.23 | BCL7B, TBL2, MLXIPL | rs17145738 | 0.13 | 7.00E-22 | |
| Kathiresan | Triglycerides | 1p31.3 | ANGPTL3, DOCK7, ATG4C | rs12130333 | 0.78 | 2.00E-08 | |
| Kathiresan | Triglycerides | 8q24.13 | TRIB1 | rs17321515 | 0.49 | 4.00E-17 | |
| Kathiresan | Triglycerides | 19p13.11 | CILP2, PBX4 | rs16996148 | 0.9 | 4.00E-09 | |
| Kathiresan | Triglycerides | 1q42.13 | GALNT2 | rs4846914 | 0.4 | 7.00E-15 | |
| Kooner | Triglycerides | 7q11.23 | MLXIPL | rs3812316 | 0.95 | 1.00E-10 | (Kooner, J.S., Chambers, J.C. *et al.*, 2008) |
| Kooner | Triglycerides | 11q23.3 | LOC440069, MGC13125 | rs1558861 | 0.18 | 2.00E-26 | |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Willer | HDL cholesterol | 12q24.11 | MVK,MMAB | rs2338104 | 0.45 | 3.00E-08 | (Willer, C.J., Sanna, S. *et al.*, 2008) |
| Willer | LDL cholesterol | 1p13.3 | CELSR2,PSRC1,SORT1 | rs599839 | 0.77 | 6.00E-33 | |
| Willer | LDL cholesterol | 6p21.32 | B3GALT4 | rs2254287 | 0.38 | 5.00E-08 | |
| Willer | Triglycerides | 2p23.3 | GCKR | rs780094 | 0.39 | 6.00E-32 | |
| Willer | Triglycerides | 1p31.3 | ANGPTL3 | rs1748195 | 0.7 | 2.00E-10 | |
| Willer | Triglycerides | 19p13.3 | NCAN,CILP2 | rs16996148 | 0.92 | 3.00E-09 | |
| Willer | Triglycerides | 8q24.13 | TRIB1 | rs17321515 | 0.56 | 7.00E-13 | |
| Willer | Triglycerides | 7q11.23 | MLXIPL | rs17145738 | 0.84 | 2.00E-12 | |
| Wallace | LDL cholesterol | 1p13.3 | CELSR2,PSRC1 | rs599839 | 0.24 | 1.00E-07 | (Wallace, C., Newhouse, S.J. *et al.*, 2008) |
| Broderick | Colorectal cancer | 18q21.1 | SMAD7 | rs4939827 | 0.52 | 1.00E-12 | (Broderick, P., Carvajal-Carmona, L. *et al.*, 2007) |
| Fox | BMI | 7q32.3 | Intergenic | rs1106683 | NR | 1.00E-07 | (Fox, C.S., Heard-Costa, N. *et al.*, 2007) |
| Fox | BMI | 7q23.3 | Intergenic | rs1106684 | NR | 2E-06 | |
| Fox | BMI | 13q21.32 | Intergenic | rs1333026 | NR | 8E-06 | |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Levy | Blood pressure | 1p31.3 | Intergenic | rs10493340 | NR | 2E-06 | (Levy, D., Larson, M.G. *et al.*, 2007) |
| Levy | Blood pressure | 8q13.3 | Intergenic | rs1963982 | NR | 0.000003 | |
| Levy | Blood pressure | 14q24.3 | Intergenic | rs935334 | NR | 0.000003 | |
| Murabito | Breast cancer | 12q21.1 | Intergenic | rs1154865 | NR | 7.00E-07 | (Murabito, J.M., Rosenberg, C.L. *et al.*, 2007) |
| Murabito | Breast cancer | 5q34 | Intergenic | rs6556756 | NR | 5.00E-07 | |
| Murabito | Breast cancer | 17q21.33 | COL1A1 | rs2075555 | NR | 8.00E-08 | |
| Murabito | Breast cancer | 18q21.2 | Intergenic | rs1978503 | NR | 1.00E-06 | |
| Murabito | Breast cancer | 13q32.1 | ABCC4 | rs1926657 | NR | 2E-06 | |
| Murabito | Prostate cancer | 3p22.2 | CTDSPL | rs9311171 | NR | 2E-06 | |
| Murabito | Prostate cancer | 13q33.1 | Intergenic | rs1529276 | NR | 2E-06 | |
| Murabito | Prostate cancer | 6p12.2 | PKHD1 | rs10498792 | NR | 0.000003 | |
| Murabito | Prostate cancer | 5q14.3 | HAPLN1 | rs4466137 | NR | 0.000003 | |
| Murabito | Prostate cancer | 3q24 | Intergenic | rs345013 | NR | 0.000005 | |
| Hakonarson | Type 1 diabetes | 16p13.13 | KIAA0350 | rs2903692 | 0.62 | 7.00E-11 | (Hakonarson, H., Grant, S.F. *et al.*, 2007) |
| Tomlinson | Colorectal cancer | 8q24.21 | Intergenic | rs6983267 | 0.49 | 1.00E-14 | (Tomlinson, I., Webb, E. *et al.*, 2007) |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Zanke | Colorectal cancer | 8q24.21 | ORF DQ515897 | rs10505477 | 0.5 | 3.00E-11 | (Zanke, B.W., Greenwood, C.M. *et al.*, 2007) |
| Gudmundsson | Prostate cancer | 17q12 | TCF2 | rs4430796 | 0.49 | 1.00E-11 | (Gudmundsson, J., Sulem, P. *et al.*, 2007b) |
| Gudmundsson | Prostate cancer | 17q24.3 | Intergenic | rs1859962 | 0.46 | 3.00E-10 | |
| WTCCC | Hypertension | 1q43 | RYR2,CHRM3,ZP4 | rs2820037 | 0.14 | 8.00E-07 | (2007) |
| WTCCC | Hypertension | 15q26 | NR | rs2398162 | 0.26 | 0.000006 | |
| WTCCC | Type 1 diabetes | 12q13.2 | ERBB3 | rs11171739 | 0.42 | 1.00E-11 | |
| WTCCC | Type 1 diabetes | 12q24.13 | SH2B3,LNK,TRAFD1,PTPN1 | rs17696736 | 0.42 | 2.00E-14 | |
| WTCCC | Type 1 diabetes | 16p13.13 | KIAA0350 | rs12708716 | 0.65 | 5.00E-07 | |
| WTCCC | Type 1 diabetes | 12p13 | NR | rs11052552 | 0.49 | 7.00E-07 | |
| WTCCC | Type 1 diabetes | 4q27 | NR | rs17388568 | 0.26 | 0.000003 | |
| WTCCC | Type 2 diabetes | 6p22.3 | CDKAL1 | rs9465871 | 0.18 | 3.00E-07 | |
| WTCCC | Type 2 diabetes | 16q12.2 | FTO | rs9939609 | 0.4 | 2.00E-07 | |
| WTCCC | Type 2 diabetes | 3p14 | NR | rs358806 | 0.8 | 0.000003 | |
| WTCCC | Type 2 diabetes | 12q15 | NR | rs1495377 | 0.5 | 7E-06 | |
| WTCCC | Type 2 diabetes | 12q13 | NR | rs12304921 | 0.15 | 7E-06 | |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Todd | Type 1 diabetes | 12q24.13 | C12orf30 | rs17696736 | 0.42 | 2.00E-16 | (Todd, J.A., Walker, N.M. *et al.*, 2007) |
| Todd | Type 1 diabetes | 12q13.2 | ERBB3 | rs2292239 | 0.34 | 2.00E-20 | |
| Todd | Type 1 diabetes | 16p13.13 | KIAA0350 | rs12708716 | 0.68 | 3.00E-18 | |
| Todd | Type 1 diabetes | 18p11.21 | PTPN2 | rs2542151 | 0.16 | 1.00E-14 | |
| Todd | Type 1 diabetes | 18q22.2 | CD226 | rs763361 | 0.47 | 1.00E-08 | |
| Easton | Breast cancer | 10q26.13 | FGFR2 | rs2981582 | 0.38 | 2.00E-76 | (Easton, D.F., Pooley, K.A. *et al.*, 2007) |
| Easton | Breast cancer | 11p15.5 | LSP1 | rs3817198 | 0.3 | 3.00E-09 | |
| Easton | Breast cancer | 8q24.21 | Intergenic | rs13281615 | 0.4 | 5.00E-12 | |
| Easton | Breast cancer | 16q12.1 | TNCR9,LOC643714 | rs3803662 | 0.25 | 1.00E-36 | |
| Easton | Breast cancer | 5q11.2 | MAP3K1 | rs889312 | 0.28 | 7.00E-20 | |
| Hunter | Breast cancer | 10q26.13 | FGFR2 | rs1219648 | 0.4 | 1.00E-10 | (Hunter, D.J., Kraft, P. *et al.*, 2007) |
| Stacey | Breast cancer | 2q35 | Intergenic | rs13387042 | 0.5 | 1.00E-13 | (Stacey, S.N., Manolescu, A. *et al.*, 2007) |
| Stacey | Breast cancer | 16q12.1 | TNRC9 | rs3803662 | 0.27 | 6.00E-19 | |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Saxena | Type 2 diabetes | 9p21.3 | CDKN2A,CDKN2B | rs10811661 | 0.83 | 8.00E-15 | (Saxena, R., Voight, B.F. *et al.*, 2007) |
| Saxena | Type 2 diabetes | 3q27.2 | IGF2BP2 | rs4402960 | 0.29 | 9.00E-16 | |
| Saxena | Type 2 diabetes | 6p22.3 | CDKAL1 | rs7754840 | 0.31 | 4.00E-11 | |
| Scott | Type 2 diabetes | 3q27.2 | IGF2BP2 | rs4402960 | 0.3 | 9.00E-16 | (Scott, L.J., Mohlke, K.L. *et al.*, 2007) |
| Scott | Type 2 diabetes | 6p22.3 | CDKAL1 | rs7754840 | 0.36 | 4.00E-11 | |
| Scott | Type 2 diabetes | 9p21.3 | CDKN2A, CDKN2B | rs10811661 | 0.85 | 8.00E-15 | |
| Scott | Type 2 diabetes | 11p12 | Intergenic | rs9300039 | 0.89 | 4.00E-07 | |
| Steinthorsdottir | Type 2 diabetes | 6p22.3 | CDKAL1 | rs7756992 | 0.26 | 8.00E-09 | (Steinthorsdottir, V., Thorleifsson, G. *et al.*, 2007) |
| Zeggini | Type 2 diabetes | 16q12.2 | FTO | rs8050136 | 0.6 | 1.00E-12 | (Zeggini, E., Weedon, M.N. *et al.*, 2007) |
| Zeggini | Type 2 diabetes | 3q27.2 | IGF2BP2 | rs4402960 | 0.32 | 9.00E-16 | |
| Zeggini | Type 2 diabetes | 9p21.3 | CDKN2A/B | rs10811661 | 0.83 | 8.00E-15 | |
| Zeggini | Type 2 diabetes | 6p22.3 | CDKAL1 | rs10946398 | 0.32 | 4.00E-11 | |
| Zeggini | Type 2 diabetes | 10q23.33 | HHEX | rs5015480 | 0.43 | 6.00E-10 | |
| Frayling | Body mass index | 16q12.2 | FTO | rs9939609 | 0.39 | 2.00E-20 | (Frayling, T.M., Timpson, N.J. *et al.*, 2007) |

**Table A.1** Full SNPs and GWAS content pack (continue).

| First Author | Disease/Trait | Region | Reported Gene(s) | SNPs | Risk Allele Frequency in Controls | p-Value | Reference |
|---|---|---|---|---|---|---|---|
| Gudmundsson | Prostate cancer | 8q24.21 | Intergenic | | 0.02 (EA) | 3.00E-15 | (Gudmundsson, J., Sulem, P. *et al.*, 2007a) |
| Gudmundsson | Prostate cancer | 8q24.21 | Intergenic | rs16901979 | 0.03 (EA) | 1.00E-12 | |
| Yeager | Prostate cancer | 8q24.21 | Intergenic | rs6983267 | 0.5 | 9.00E-13 | (Yeager, M., Orr, N. *et al.*, 2007) |
| Sladek | Type 2 diabetes | 10q23.33 | HHEX | rs1111875 | 0.4 | 0.000003 | (Sladek, R., Rocheleau, G. *et al.*, 2007) |
| Sladek | Type 2 diabetes | 8q24.11 | SLC30A8 | rs13266634 | 0.3 | 6.00E-08 | |

# CURRICULUM VITAE (II)

| | |
|---|---|
| **NAME** | Mr. Palang Chotsiri |
| **DATE OF BIRTH** | 1 December 1984 |
| **EDUCATION RECORD** | |
| HIGH SCHOOL | High School Graduation<br>The Laboratory School of Rajabhat Institute Phranakhon Si Ayutthaya, 2003 |
| BACHELOR'S DEGREE | Bachelor of Science (Physics)<br>Mahidol University, 2007 |
| MASTER'S DEGREE | Master of Science (Bioinformatics)<br>King Mongkut's University of Technology Thonburi, 2011 |
| **SCHOLARSHIP** | Full Scholarship, by National Center for Genetic Enginnering and Biotechnology and King Mongkut's University of Technology Thonburi for Master's Degree in Bioinformatics, 2007<br><br>Thailand Full Scholarship for Distinguish Science Student, Ministry of Science and Technology of Thailand for Bachelor's degree in Sciences, 2004 |
| **PUBLICATION** | Chotsiri P., Cheevadhanarak S., Senachak J., Laoteng K., Paithoonrangsarid K., Plengvidhaya V., 2008, "Evolutionary scenarios of cyanobacterial lineage determining by comparative genomics". **International conference on life science 2008 (BioAsia 2008)**, Bangkok, Thailand, (poster presentation)<br><br>Saehuan C., Chotsiri P., Rojanarata T., 2008, "Molecular phylogeny based on benzoylformate decarboxylase of Pseudomonas strains", **ก้าวทันโลกวิทยาศาสตร์ ปีที่ 8(1)**, 78-76.<br><br>Chosiri P., Plengvidhaya V., Senachak J., Paithoonrangsarid K., Laoteng K., Prommeenate P., Cheevadhanarak S., 2009, "Uncovering photosynthesis apparatus and genomic repertoires of cyanobacterial ancestor via phylogenomic analysis". **Thailand Society of Biotechnology (TSB)**. Bangkok, Thailand, (oral presentation) |

**King Mongkut's University of Technology Thonburi**

**Agreement on Intellectual Property Rights Transfer for Postgraduate Students**

Date.....

Name.....Mr. Palang.....Middle Name.....

Surname/Family Name.....Chotsiri.....

Student Number....50460004..... who is a student of King's Mongkut's University of Technology Thonburi (KMUTT) in ❍ Graduate Diploma ☑ Master Degree ❍ Doctoral Degree

Program.....Bioinformatics.....Field of Study..Bioinformatics.....

Faculty/School..School of Bioresources and Technology and School of Information Technology

Home Address ....57 Moo. 4, Bang-ra-kham.....

.....Nakhon-laung, Ayudhaya.....

Postal Code...13260.....Country.....Thailand.....

I, as 'Transferer', hereby transfer the ownership of my thesis copyright to King's Mongkut's University of Technology Thonburi and National Center for Genetic Engineering and Biotechnology, Thailand who has appointed (Dean's name) Assoc.Prof. Narumon Jeyashoke and Assoc.Prof.Dr. Nipon Charoenkitkarn, Dean of School of Bioresources and Technology and Dean of School of Information Technology to be 'Transferees' of copyright ownership under the 'Agreement' as follows.

1. I am the author of the thesis entitled ..."Evolutionary scenarios of cyanocaterial lineate as determined by comparative genomics approach (I)" and "Single nucleotide Polymorphisms (SNPs) and genome wide association study (GWAS) content pack (II)"... under the supervision of ...Dr. Vethachai Plengvidhya..... who is my supervisor, and/or.Assoc. Prof. Dr. Supapon.Cheevahanarak and Dr. Jittisak Senachak. who is/are my co-supervisor(s), in accordance with the Thai Copyright Act B.E. 2537. The thesis is a part of the curriculum of KMUTT.

2. I hereby transfer the copyright ownership of all my works in the thesis to KMUTT throughout the copyright protection period in accordance with the Thai Copyright Act B.E. 2537, effective on the approval date of thesis proposal consented by KMUTT.

3. To have the thesis distributed in any form of media, I shall in each and every case stipulate the thesis as the work of KMUTT.

4. For my own distribution of thesis or the reproduction, adjustment, or distribution of thesis by the third party in accordance with the Thai Copyright Act B.E. 2537 with remuneration in return, I am subject to obtain a prior written permission from KMUTT.

5. To use any information from my thesis to make an invention or create any intellectual property works within ten (10) years from the date of signing this Agreement, I am subject to obtain prior written permission from KMUTT, and KMUTT is entitled to have intellectual property

rights on such inventions or intellectual property works, including entitling to take royalty from licensing together with the distribution of any benefit deriving partly or wholly from the works in the future, conforming with the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.

6. If the benefits arise from my thesis or my intellectual property works owned by KMUTT, I shall be entitled to gain the benefits according to the allocation rate stated in the Regulation of King Mongkut's Institute of Technology Thonburi *Re* the Administration of Benefits deriving from Intellectual Property B.E. 2538.

Signature.....Transferor

(Mr. Palang Chotsiri)

Student

Signature.....Transferee

(Assoc.Prof. Narumon Jeyashoke)

Dean

Signature.....Transferee

(Assoc.Prof.Dr. Nipon Charoenkitkarn)

Dean

Signature.....Witness

(Dr. Vethachai Plengvidhya)

Signature.....Witness

(Asst.Prof.Dr. Marasri Ruengjitchatchawalya)