Journal Club – Bioinformatics and Biostatistics

Han Fang
hfang@cshl.edu
Lyon Lab

Contents

- Whole Genome Sequencing in support of Wellness and Health Maintenance
- Likelihood ratios for genome medicine
- RGLM=random generalized linear model
- Possible pipeline utilizing the above methods
- Family-based association studies on NGS data
- An Efficient Sufficient Dimension Reduction Method for Identifying Genetic Variants of Clinical Significance

Whole Genome Sequencing in support of Wellness and Health Maintenance

Genome Medicine 2013, 5:58 doi:10.1186/gm462
Chirag J Patel et.al

Overall idea

Here we begin to explore how whole genome sequencing (WGS) might be incorporated alongside traditional clinical evaluation as a part of preventive medicine. The present study illustrates novel approaches for integrating genotypic and clinical information for assessment of generalized health risks and to assist individuals in the promotion of wellness and maintenance of good health.

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

diverse as diabetes, asthma and depression [9, 10]. Since environment makes a substantial contribution to these latter conditions, prediction is too strong a claim for genomic medicine [11, 12], but risk stratification is certainly feasible [13]. Here we explore how whole genome

Feero WG, Guttmacher AE, Collins FS: Genomic medicine - an updated primer. N Engl J Med 2010, 362:2001-2011.

Methods

- WGS & longitudinal clinical profiles of 8 ppl (4+4)
- Multivariate genotypic risk assessments (GWAS) I.E. Common variants
- Clinical measurements : immune, metabolic, cardiovascular, musculoskeletal, respiratory, and mental health.

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

Methods - Cont.

- Non-random picking of samples
- Clinical Assessments

1) Health status assessments of interest,

2) Self-reported family and personal medical history

- WGS

1) Mean coverage= ~36X with 95.5% (mean) >10X

2) 87% of each individual's quality filtered reads were aligned

3) HiSeq2000

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

Methods - Cont.

- Genetic risk assessment (common variants)

1) Genetic risk predictions from VARIMED DB
2) Combine odds ratios of associated SNPs with diseases and traits (LR Paper)

$$\log LR(x) = \frac{\sum_{i=1}^{s} \log \frac{F(g\ in\ cases)}{F(g\ in\ controls)} \times \sqrt{S(i)}}{\sum_{i=1}^{s} \sqrt{S(i)}}$$

3) Loci with P-value < 10^{-6} & site in a haplotype block (r^2 >= 0.8)

$$\text{Post-test probability} = \text{Pretest Odds} * \prod_{i=1}^{n} LR(i)$$ Wrong

Post test odds

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

LR paper

Calculation of likelihood ratios, and pre- and post-test probabilities

Likelihood ratio = Probability of genotype in diseased person/Probability of genotype in non-diseased person
Likelihood ratios multiplied by the pre-test odds of disease give the post-test odds of disease (Table 1), and these likelihood ratios may be chained together (Figure 1):

Pre-test odds = Probability of disease/1 - Probability of disease
Pre-test odds * LR1 * LR2 ….. *LRn = Post-test odds

Post-test probability = Post-test odds/Post-test odds + 1

Table 1. Example calculations of post-test probabilities

Type of disease and associated variants probability of disease (%)	Pre-test probability of disease (%)	Likelihood ratio	Post-test
Common disease, weakly associated variant	15.0	1.1	16.256
Common disease, several weakly associated variants	15.0	1.1 × 1.1 × 1.1 × 1.1 = 1.46	20.486
Rare disease, weakly associated variant	0.01	1.1	0.011
Rare disease, strongly associated variant	0.01	5	0.050
Rare disease, several weakly associated variants	0.01	1.1 × 1.1 × 1.1 × 1.1 = 1.46	0.015
Rare disease, several moderately associated variants	0.01	2 × 2 × 2 × 2 = 16	0.160

Post-test probabilities may be calculated for common or rare diseases with weakly and strongly associated variants using example values for likelihood ratios and pre-test probabilities. The definition of strongly versus weakly associated is in the context of genetic associations, where likelihood ratios from large-scale studies rarely reach higher than 3. Many clinical laboratory tests have likelihood ratios of 10 or more.

Morgan AA, Chen R, Butte AJ: Likelihood ratios for genome medicine. Genome Med 2010, 2:30.

LR paper - Cont.

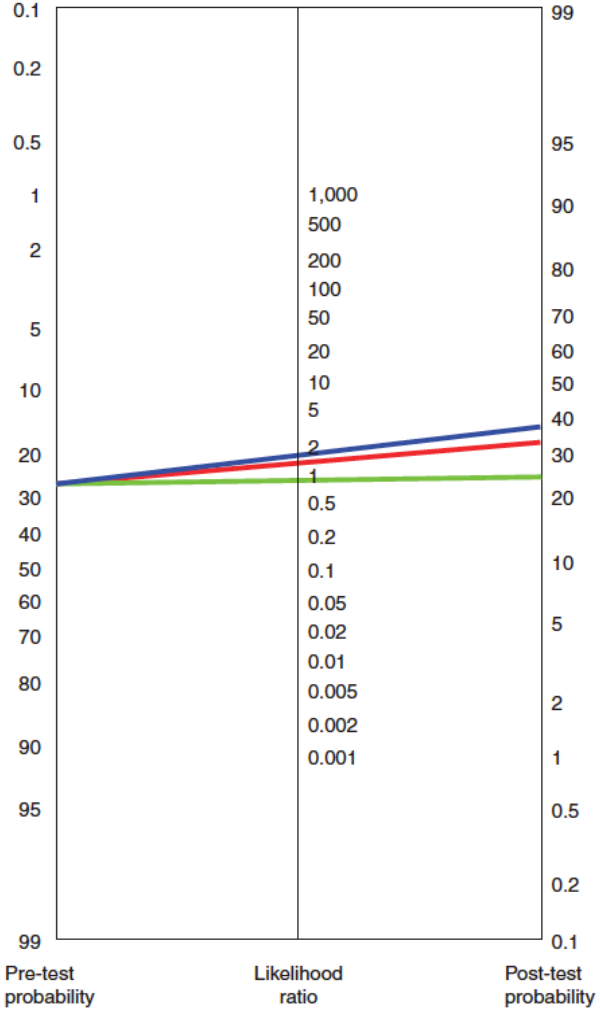


Figure 1. Nomogram for likelihood ratios. The pre-test and post-test probabilities and likelihood ratios of any diagnostic test, including a genetic test, can be visualized using a nomogram familiar to most physicians and medical students. The nomogram shown is derived from the Fagan nomogram [14], and modified from one generated using a web-based tool [28]. The left side of the figure indicates a hypothetical pre-test probability of disease of 27%. Three lines represent the three possible genotypes, from top to bottom: homozygous risk alleles with a likelihood ratio of 1.61, heterozygous alleles with a likelihood ratio of 1.26, and homozygous protective alleles with a likelihood ratio of 0.83. The right side of the figure indicates three possible post-test probabilities resulting from the three genotypes. Multiple such tests can be 'chained' together serially, if they describe independent risks and cover the same pre-test assumptions.

Morgan AA, Chen R, Butte AJ: Likelihood ratios for genome medicine. Genome Med 2010, 2:30.

LR Paper – Interesting quotes

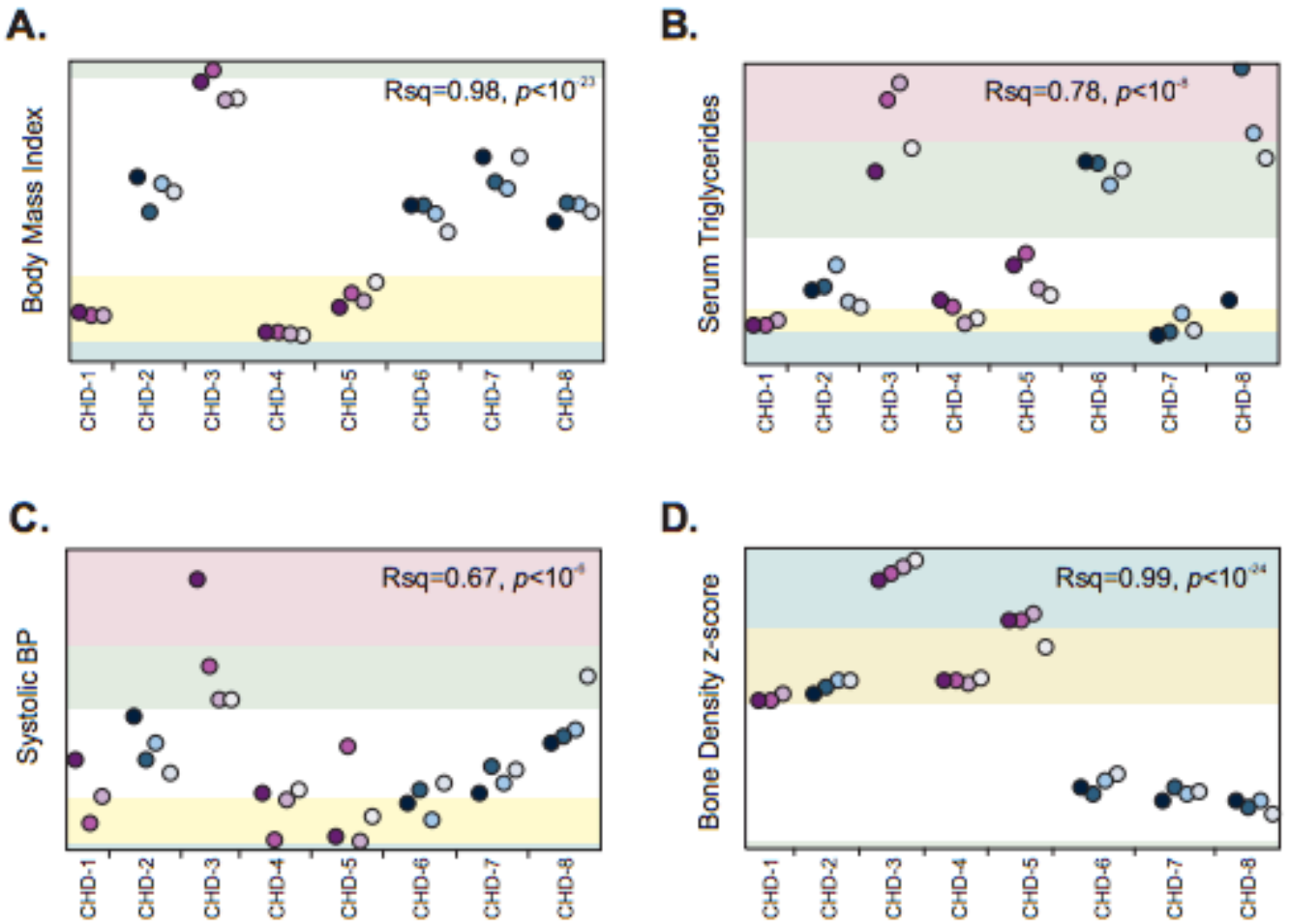
Traditionally, the published literature on genetic associations has focused on suggesting interesting variants with possible mechanistic involvement in the disease of study. Hence, authors may only report an odds ratio as a measure of effect size, and a *P* value to show that the variant is significantly associated with the disease. Many such studies do not even report the risk genotype at the site of the SNP; this is a particular problem because the relationship of the common allele in the population under study to a reference genome is unknown, and the reference genome may actually contain the risk-associated allele.

We recently curated 2,174 articles reporting primary data on gene-disease associations of variants in the National Center for Biotechnology Information (NCBI) SNP database (dbSNP) [20]. Of these publications, only 46% contained information on actual genotype-associated risk, enabling the calculation of a likelihood ratio yielding a total of 2,092 disease-variant associati . Although any particular genetic association study may not be intended for use in informing a clinical diagnostic test or interpretation, information on the actual proportion/frequency of subjects with each associated genotypic variant in the relevant phenotype categories (such as with and without disease) should be made available for use in further studies and meta-analyses. This information aids in attempts at replication of results and in calculating overall estimates of the power of a particular genotype to predict disease state.

Morgan AA, Chen R, Butte AJ: Likelihood ratios for genome medicine. Genome Med 2010, 2:30.

Methods – Cont. Clinical risk assessment

- 1) 5 levels of risks regarding 8 different disease categories



CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

Methods – Integration of genetic and clinical data

Directly matches GWAS results with individual diseases

- average Framingham risk scores (FRS) for each person at each visit across the entire CHDWB database
- post-test probability=pre-test probability * LR/(1+(pre-test probability*(LR-1)))

- Limitations:
- no appropriate clinical biomarkers for some diseases in our cohort
- Some are precisely the endophenotype of the disease/ traits investigated in GWAS (Redundant)

Combines multiple clinical and genetic measures

- generate an overall portrait of risk in eight major disease categories
- The z-scores for clinical parameters are adjusted with respect to risk predisposition: for traits which are known to confer risk at a lower level
- genetic risk scores are ranked according to percentiles into five categories

Results - Cont.

Table 1: Summary of variations in genome sequences of eight Caucasian subjects

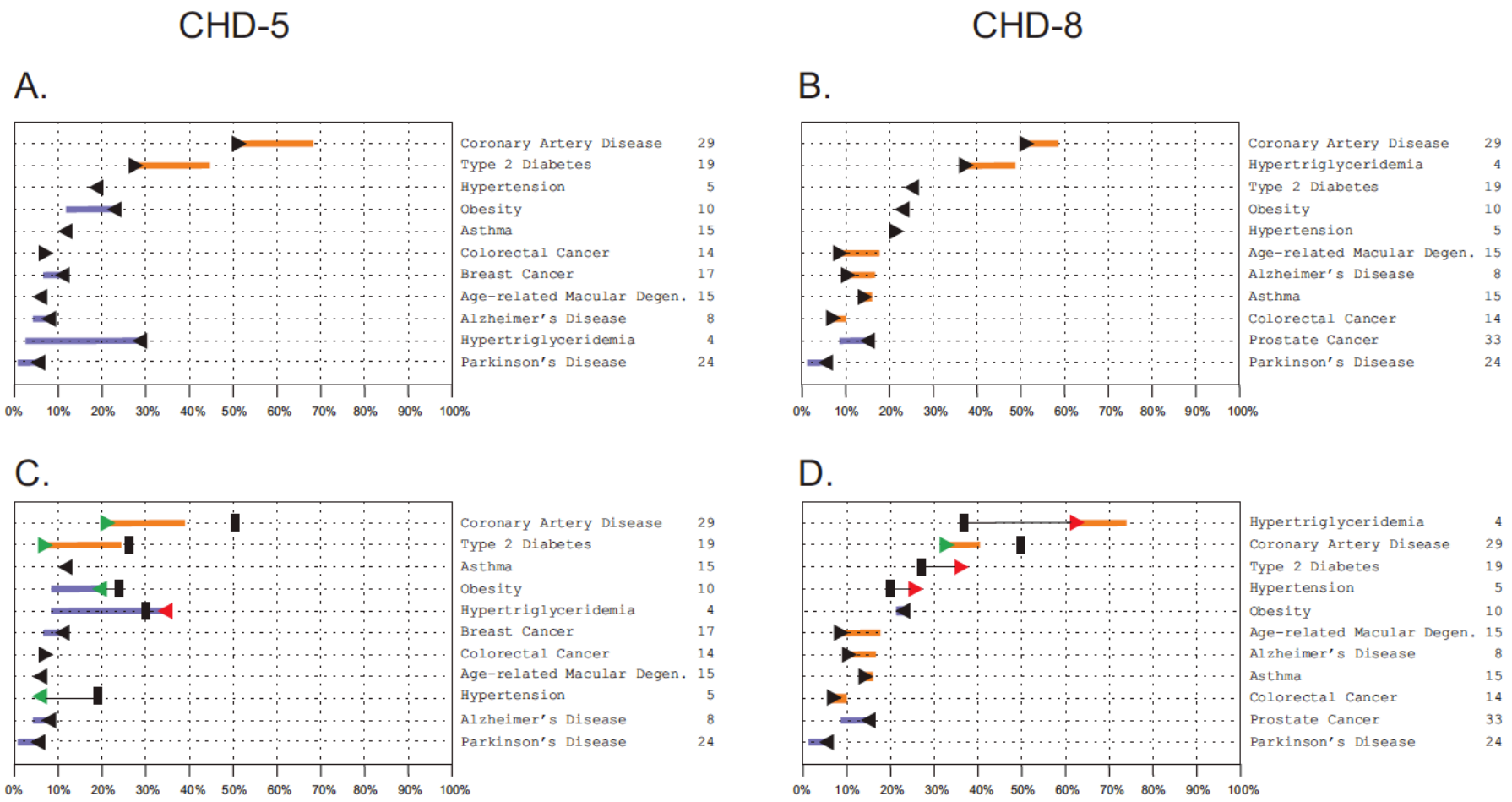
SampleID	Total Variants (>q20)			Coding variants							
	SNPs	Indels	SVs	SNP				Indel			SV
				Synonymous (rare homo)	Missense (rare homo)	Nonsense	Splice Overlap	Indel FS	Indel NFS	Indel Overlap	
CHD 1	3722234	641792	4197	11887 (18)	11434 (29)	64	76	303	299	125	41
CHD 2	3701558	639005	4739	11842 (11)	11708 (33)	60	81	334	290	118	37
CHD 3	3691270	632544	4033	11912 (9)	11488 (31)	65	71	279	304	116	37
CHD 4	3691337	633475	4114	11757 (9)	11457 (25)	56	90	317	280	106	49
CHD 5	3734820	645032	3977	11929 (9)	11745 (35)	62	90	343	307	123	43
CHD 6	3650690	602744	3916	11560 (12)	11285 (37)	60	80	342	280	112	32
CHD 7	3643046	597363	4011	11814 (17)	11480 (41)	61	85	289	287	109	31
CHD 8	3647944	590064	3828	11619 (9)	11255 (18)	54	76	311	281	95	38
Pelak et al.[27]	*3473639*	*609795*	*805 (CNVs)*	-	*11069*	*117*	*99*	*479*	*898*	-	-
Shen et al.[40]	*3307678*	*421088*	-	*9612*	*9082*	*87*	-	*217*	*164*	-	-

SV: structural variants; FS: frameshift; NFS: non frameshift; overlap: located within 2nt of exon-intron boundary.

The coding variants were classified based on Gencode v7

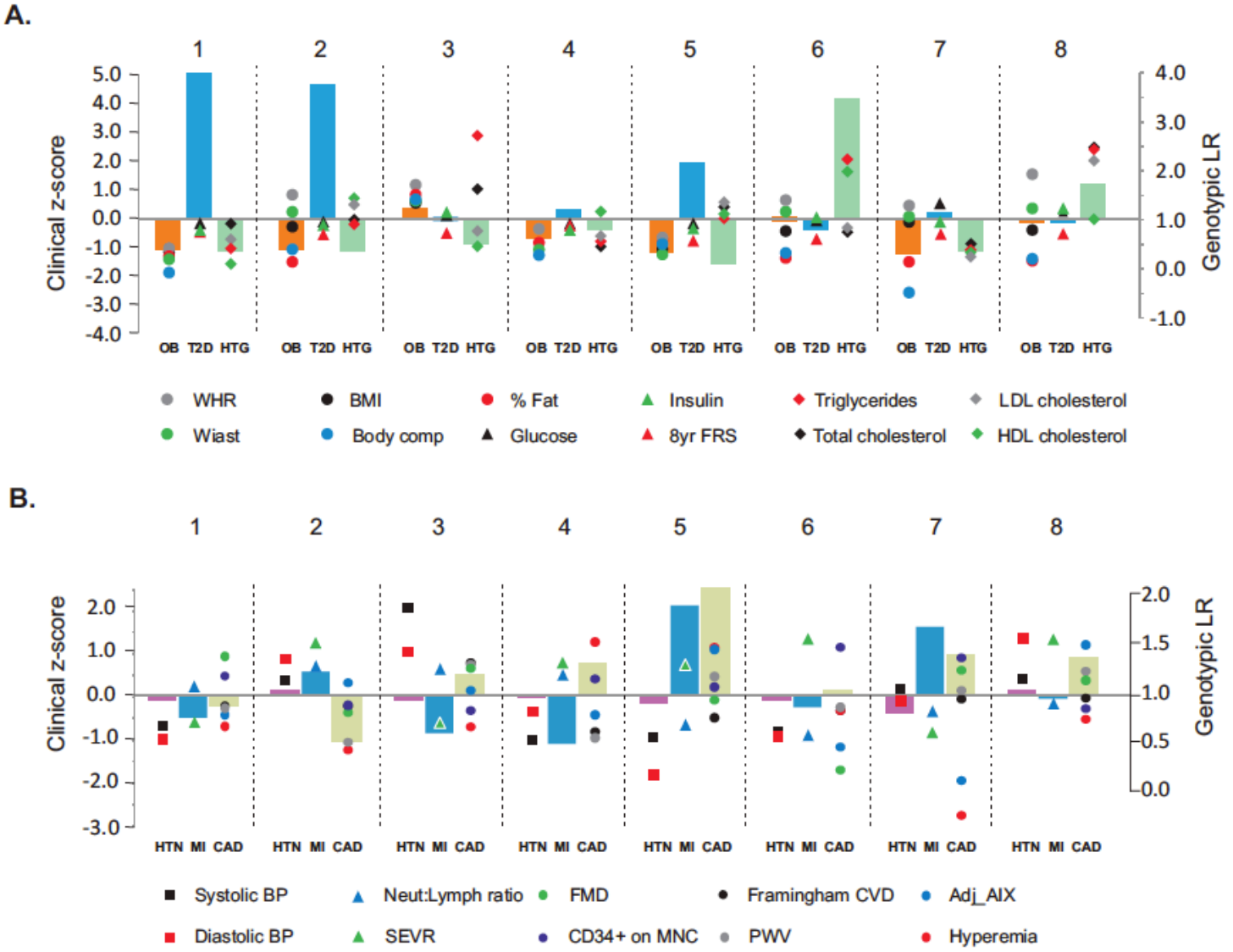
CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

Results – Method1



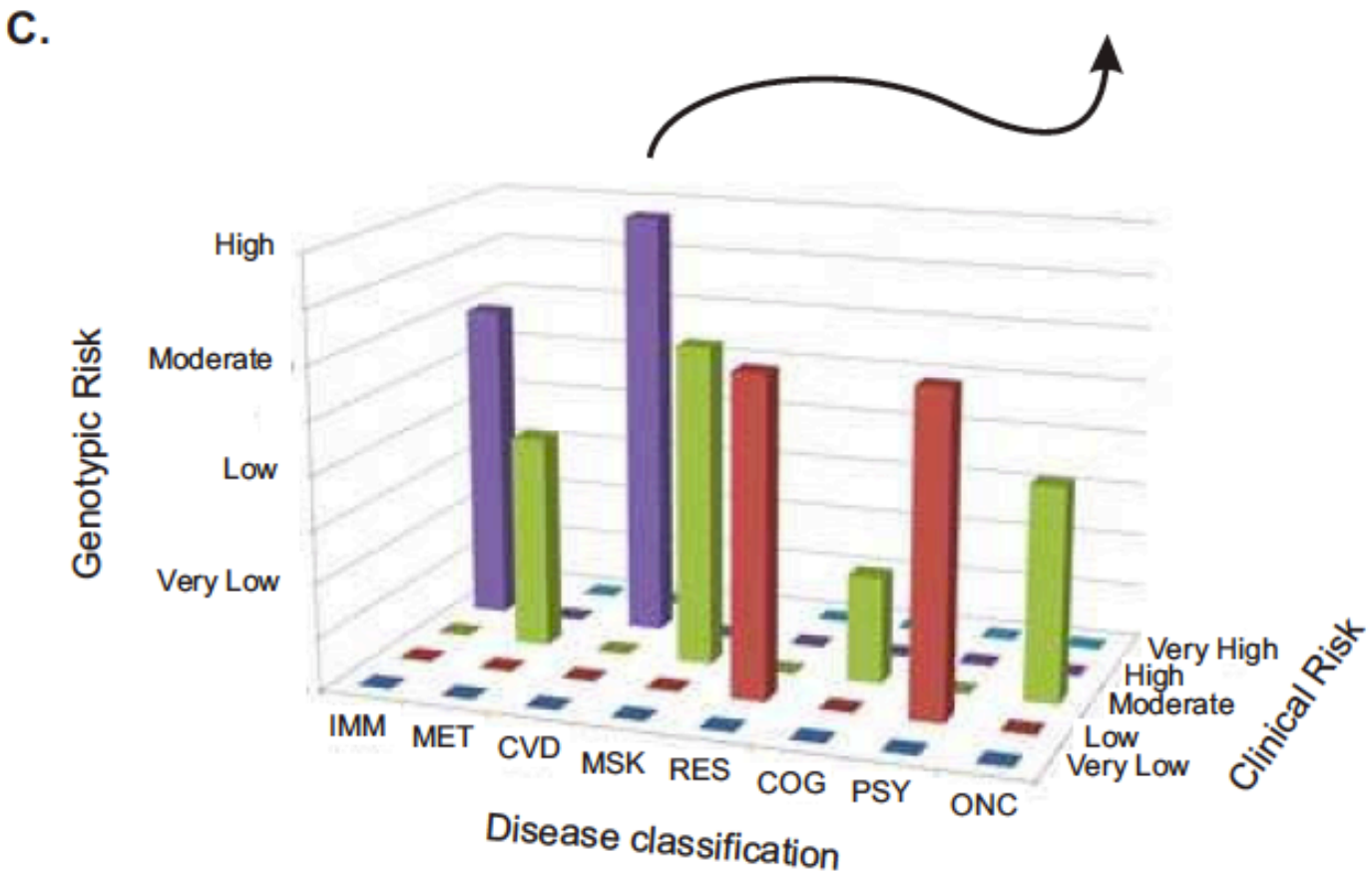
(A) (B) standard format; (C) (D) risk-o-grams adjusted according to observed clinical risk. Gender and age-specific disease prevalence for Caucasians are indicated by the black triangles (or hash marks in lower panels). Genotypic effects are predicted to increase (right point) or decrease (left point) overall risk by the indicated magnitude (orange or purple, respectively), resulting in the indicated rank-ordered overall risk. The number of SNPs used in the computation for each disease is indicated to the right. The adjusted risk-o-grams according to observed clinical risk (C and D) either increases (red triangle) or decreases (green triangle) the baseline without affecting the genotypic component, but results in adjustment of overall risk and rank order. Note that CHD-5 is a woman, and CHD-8 a man, so breast and prostate cancer are indicated for each as appropriate.

Method 2



(A) Metabolic risk showing joint genotypic likelihood ratios for Obesity (OB), Type 2 Diabetes (T2D), and Hypertriglyceridemia (HTG) as bars, and related clinical measures as the indicated points, for each of the 8 individuals in the study. (B) Similar Cardiovascular risk assessments for Hypertension (HTN), Myocardial Infarction (MI), and Coronary Artery Disease (CAD) as bars, along with related clinical measures and/or Framingham Risk Score. All measures are averaged over the first three visits to the CHDWB.

Results – Method2



(C) Proposal for "Gridiron Plot" representation of clinical risk (y-axis) against genotypic risk (z-axis) in eight disease domains described in the text, for individual CHD-5. The plot gives a glimpse of where the two types of assessment are concordant (eg CVD, cardiovascular) or discordant (IMM, immunological), and more refined analyses such as in panels (A) and (B) provide further clues as to the genetic basis of overall risk.

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

References

CJ Patel et.al: Whole Genome Sequencing in support of Wellness and Health Maintenance. Genome Medicine 2013, 5:58 doi:10.1186

Feero WG, Guttmacher AE, Collins FS: Genomic medicine - an updated primer. N Engl J Med 2010, 362:2001-2011.

Song L, Langfelder P, Horvath S (2013) Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinformatics

Morgan AA, Chen R, Butte AJ: Likelihood ratios for genome medicine. Genome Med 2010, 2:30.

Yun Zhu and Momiao Xiong, Family-Based Association Studies for Next-Generation Sequencing, The American Journal of Human Genetics 90, 1028–1045

Li Luo, Yun Zhu and Momiao Xiong*, Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation, European Journal of Human Genetics (2013) 21, 217–224

Momiao Xiong, Long Ma, An Efficient Sufficient Dimension Reduction Method for Identifying Genetic Variants of Clinical Significance