NIH National Heart, Lung, and Blood Institute

GENETIC RISK FACTORS FOR PRETERM BIRTH

Sabah Munir, *Wartburg College*
Journey Penney, *Willamette University*
Advisor: Dr. Patrick Breheny, *University of Iowa College of Public Health*

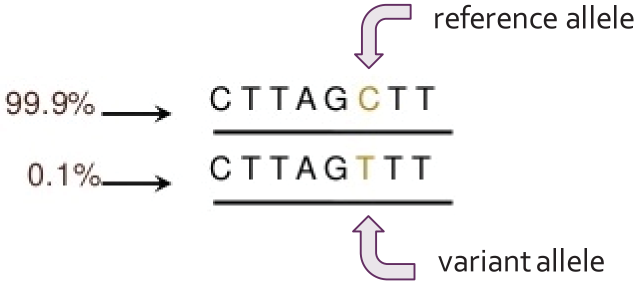
Iowa Summer Institute of Biostatistics

Background

- Preterm Birth
    - Occurs when baby is born before 37 completed weeks of gestation
        - Normally, pregnancies last around 40 weeks
    - Factors
        - Smoking, nutrition, race, age
        - Genetics
    - Affects 5-18% of pregnancies worldwide
    - Leading cause of death in children under 5 years old

Background

- Genetic Terminology:
  - **Nucleotide**
    - Building blocks of DNA. Four bases: Adenine, Cytosine, Guanine, Thymine
  - **Genetic Variant**
    - Nucleotide causing variation from most common DNA sequence



reference allele

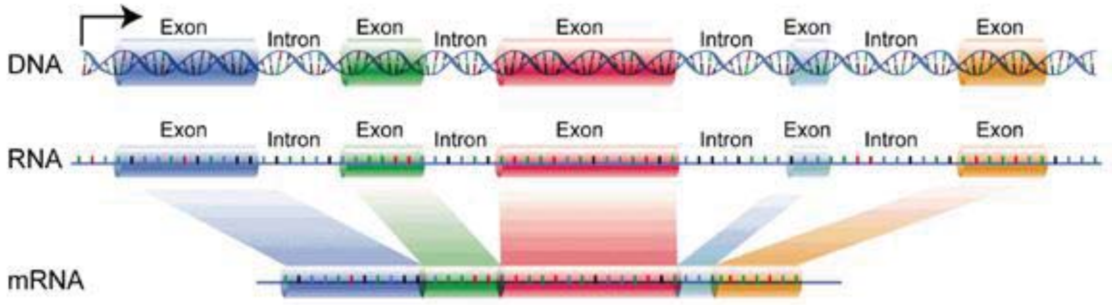
variant allele

  - **Minor Allele Frequency (MAF)**
    - The frequency of a variant allele occurring in the population
    - Rare variants: MAF < 2%

Image URL: https://www.slideshare.net/dasbipul/single-nucleotide-polymorphism-41650224

Background

- **Exon**
  - Coding region of a gene
  - The portion that is ultimately expressed as protein (DNA->RNA->Protein)
  - **Exome:** collection of all the exons in an individual's DNA

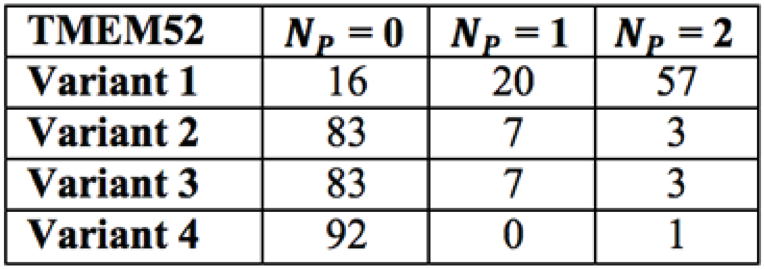


- **Whole exome sequencing:** determines nucleotide order of the exome
  - Cheaper, more practical than sequencing entire genome

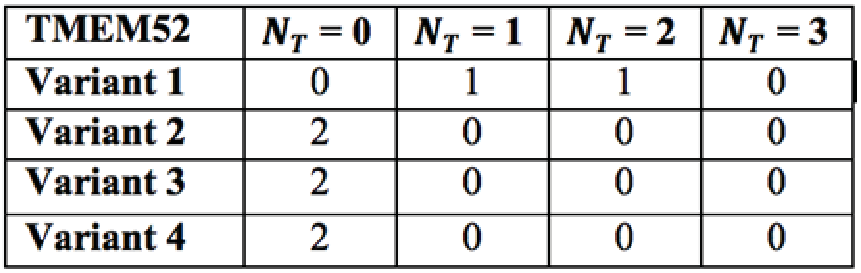
Image URL: http://de.academic.ru/pictures/dewiki/68/DNA_exons_introns.gif

Study Design

- Our Data
  - Used whole exome sequencing
  - Participants
    - Women of European ancestry (Denmark), history of preterm birth
    - 93 sister pairs, 2 sister trios (originally 97 pairs)
  - Example:



TMEM52	$N_P = 0$	$N_P = 1$	$N_P = 2$
Variant 1	16	20	57
Variant 2	83	7	3
Variant 3	83	7	3
Variant 4	92	0	1



TMEM52	$N_T = 0$	$N_T = 1$	$N_T = 2$	$N_T = 3$
Variant 1	0	1	1	0
Variant 2	2	0	0	0
Variant 3	2	0	0	0
Variant 4	2	0	0	0

Research Goals

- Develop tests to analyze PTB data against Exome Aggregation Consortium (**ExAC**) data
  - Use exome sequencing data from ExAC as general population
    - Provided us with MAF values
- Identify rare variants that influence the risk of preterm birth
- Compare two methods of statistical analysis that we developed
  - Count-based approach, treats all variants equally
  - Weighted approach, emphasizes variants with larger impact

Research Design and Methods

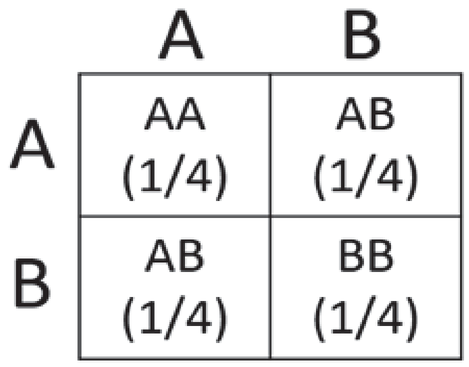
- Gene Burden Tests
  - Common way to examine whole exome sequencing
  - Combine variants on the same gene and then conduct the test
  - Test at the gene level rather than test each variant
    - 16,934 genes vs. 98,679 variants
  - Fewer tests will increase power

Research Design and Methods

- Assumptions:
  - Known: Punnett Square Probabilities
    - Shows genetic combinations possible for child
    - Can be used to find likelihoods for sibling sets

      No minor allele (AA) = ¼
      1 minor allele (AB) = ¼ + ¼ = ½
      2 minor alleles (BB) = ¼

  - Presumed: Hardy-Weinberg Equilibrium
    - Use of variables (p, q)



	A	B
A	AA (1/4)	AB (1/4)
B	AB (1/4)	BB (1/4)

Research Design and Methods

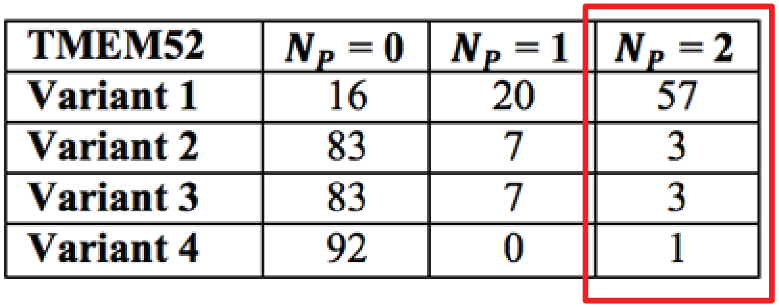
Parents	H-W Probability	Sister Pairs			Sister Trios			
		0	**1**	**2**	**0**	**1**	**2**	**3**
AA AA	q^4	1	0	0	1	0	0	0
AA AB	$4pq^3$	1/4	1/2	1/4	1/8	3/8	3/8	1/8
AB AB	$4p^2q^2$	1/16	6/16	9/16	1/64	9/64	27/64	27/64
AA BB	$2p^2q^2$	0	0	1	0	0	0	1
AB BB	$4p^3q$	0	0	1	0	0	0	1
BB BB	p^4	0	0	1	0	0	0	1

$$\mathbf{P(N_P=2)} = q^4(0) + 4pq^3(\tfrac{1}{4}) + 4p^2q^2(9/16) + 2p^2q^2(1) + 4p^3q(1) + p^4(1) = \mathbf{pq^3 + 2.25p^2q^2 + 2p^2q^2 + 4p^3q + p^4}$$

Research Design and Methods

- Count-Based Test:
  - Poisson Distribution
  - Select data included (burden test)
  - Expected Counts:

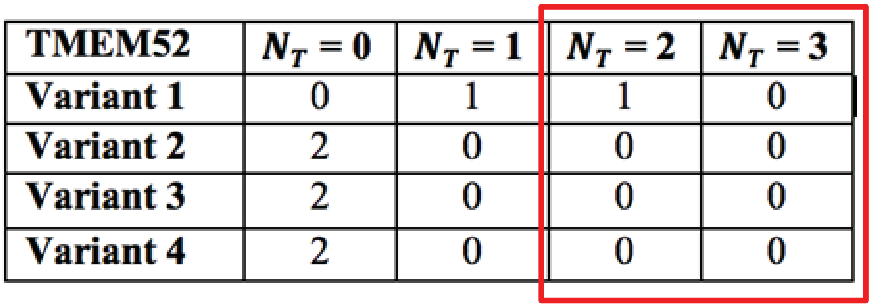
$$\sum_{i=1}^{3}\{n_T(p_{Ti3}+p_{Ti2})+n_P(p_{Pi2})\}$$



TMEM52	$N_P=0$	$N_P=1$	$N_P=2$
Variant 1	16	20	57
Variant 2	83	7	3
Variant 3	83	7	3
Variant 4	92	0	1

Observed Counts: (3+3+1)

$$\sum_{i=1}^{3}\{(\sum N_{Ti3}+N_{Ti2})+\sum N_{Pi2}\}$$



TMEM52	$N_T=0$	$N_T=1$	$N_T=2$	$N_T=3$
Variant 1	0	1	1	0
Variant 2	2	0	0	0
Variant 3	2	0	0	0
Variant 4	2	0	0	0

Research Design and Methods

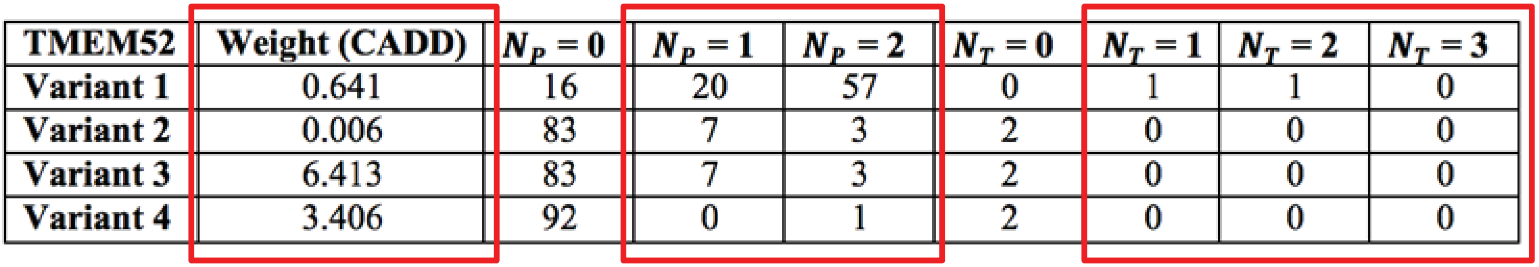
- Shortcomings of Count-Based Test:
  - All variants are treated with equal importance
    - Counts are not weighted
    - Doesn’t reflect the magnitude of “harmful” variants
  - Neglects the N=1 column

- Extending our analysis:
  - Develop a test which incorporates CADD score
    - CADD: quantifies how negatively a variant impacts the gene
    - 0-10: benign mutation; 10-20: ambiguous impact; 20+: deleterious

Research Design and Methods

- Weighted Test:
  - Normal distribution
  - Different weights for each variant, where weight = CADD score
  - Gives increasing importance to N=1, N=2, N=3

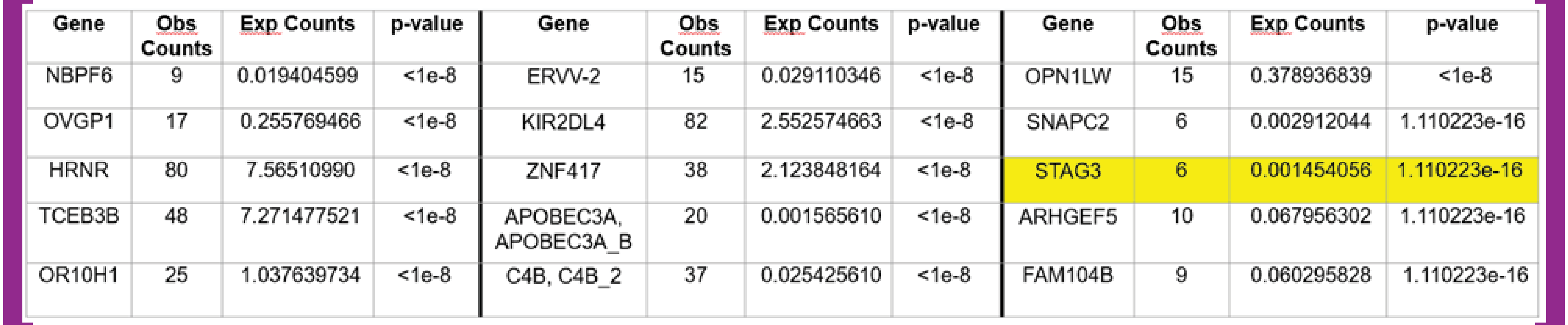
Weighted Obs. Score: $\sum_{i=1}^{4} w_i\{(2N_{Pi2} + N_{Pi1}) + \{(3N_{Ti3} + 2N_{Ti2} + N_{Ti1})\}$



TMEM52	Weight (CADD)	$N_P = 0$	$N_P = 1$	$N_P = 2$	$N_T = 0$	$N_T = 1$	$N_T = 2$	$N_T = 3$
Variant 1	0.641	16	20	57	0	1	1	0
Variant 2	0.006	83	7	3	2	0	0	0
Variant 3	6.413	83	7	3	2	0	0	0
Variant 4	3.406	92	0	1	2	0	0	0

Results and Discussion

- Top 15 Genes (Count-Based Test, p-value):



Gene	Obs Counts	Exp Counts	p-value	Gene	Obs Counts	Exp Counts	p-value	Gene	Obs Counts	Exp Counts	p-value
NBPF6	9	0.019404599	<1e-8	ERVV-2	15	0.029110346	<1e-8	OPN1LW	15	0.378936839	<1e-8
OVGP1	17	0.255769466	<1e-8	KIR2DL4	82	2.552574663	<1e-8	SNAPC2	6	0.002912044	1.110223e-16
HRNR	80	7.56510990	<1e-8	ZNF417	38	2.123848164	<1e-8	STAG3	6	0.001454056	1.110223e-16
TCEB3B	48	7.271477521	<1e-8	APOBEC3A, APOBEC3A_B	20	0.001565610	<1e-8	ARHGEF5	10	0.067956302	1.110223e-16
OR10H1	25	1.037639734	<1e-8	C4B, C4B_2	37	0.025425610	<1e-8	FAM104B	9	0.060295828	1.110223e-16

Results and Discussion

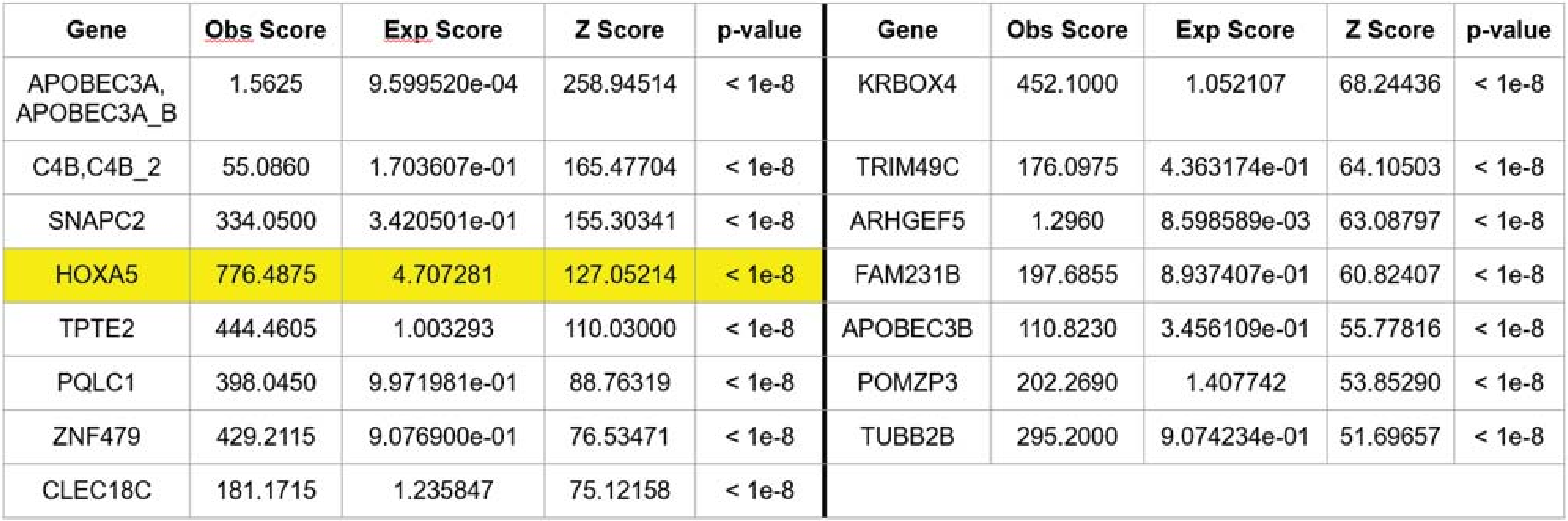
- Gene of Interest (Count-Based Test): STAG3

Gene	Obs Counts	Exp Counts	p-value
STAG3	6	0.001454056	1.110223e-16

	Ref	Alt	ExAC	CADD	P0	P1	P2	T0	T1	T2	T3
V1	G	T	0.2275	0.135	86	7	0	2	0	0	0
V2	A	C	0.2275	0.099	72	20	1	1	0	1	0
V3	A	C	0.4788	0.003	13	22	58	0	0	0	2
V4	G	A	0.000015	16.670	57	31	5	1	0	1	0
V5	T	A	0.2519	10.010	38	26	29	1	0	0	1

Results and Discussion

- Top 15 Genes (Weighted Test, z-score):



Gene	Obs Score	Exp Score	Z Score	p-value
APOBEC3A, APOBEC3A_B	1.5625	9.599520e-04	258.94514	< 1e-8
C4B,C4B_2	55.0860	1.703607e-01	165.47704	< 1e-8
SNAPC2	334.0500	3.420501e-01	155.30341	< 1e-8
HOXA5	776.4875	4.707281	127.05214	< 1e-8
TPTE2	444.4605	1.003293	110.03000	< 1e-8
PQLC1	398.0450	9.971981e-01	88.76319	< 1e-8
ZNF479	429.2115	9.076900e-01	76.53471	< 1e-8
CLEC18C	181.1715	1.235847	75.12158	< 1e-8
KRBOX4	452.1000	1.052107	68.24436	< 1e-8
TRIM49C	176.0975	4.363174e-01	64.10503	< 1e-8
ARHGEF5	1.2960	8.598589e-03	63.08797	< 1e-8
FAM231B	197.6855	8.937407e-01	60.82407	< 1e-8
APOBEC3B	110.8230	3.456109e-01	55.77816	< 1e-8
POMZP3	202.2690	1.407742	53.85290	< 1e-8
TUBB2B	295.2000	9.074234e-01	51.69657	< 1e-8

Results and Discussion

- Gene of Interest (Weighted Test): HOXA5

Gene	Obs Score	Exp Score	Z Score	p-value
HOXA5	776.4875	4.707281	127.05214	< 1e-8

	Ref	Alt	ExAC	CADD	P0	P1	P2	T0	T1	T2	T3
V1	G	A	0.000063	7.756	92	1	0	2	0	0	0
V2	C	G	0.000025	20.400	58	32	3	2	0	0	0
V3	A	G	0.0056	1.691	89	3	1	2	0	0	0

Results and Discussion

- Limitations
  - Confounding variables
  - Source of data (only one ethnic group studied)
- Replication studies
  - Follow-up study to confirm importance of rare variants found
- Refinements to weighted testing approach
  - Normal distribution not accurate for extremely rare variants
  - Weights are relative within genes, not absolute

NIH National Heart, Lung, and Blood Institute

Acknowledgments

- Dr. Patrick Breheny
- Dr. Kelli Ryckman
- Dr. Gideon Zamba
- Monica Ahrens
- ISIB Cohort Members
- National Heart, Lung, and Blood Institute (NHLBI, Grant No. HL131467)