

Statistical methods for finding miRNA motifs

Scott Norton

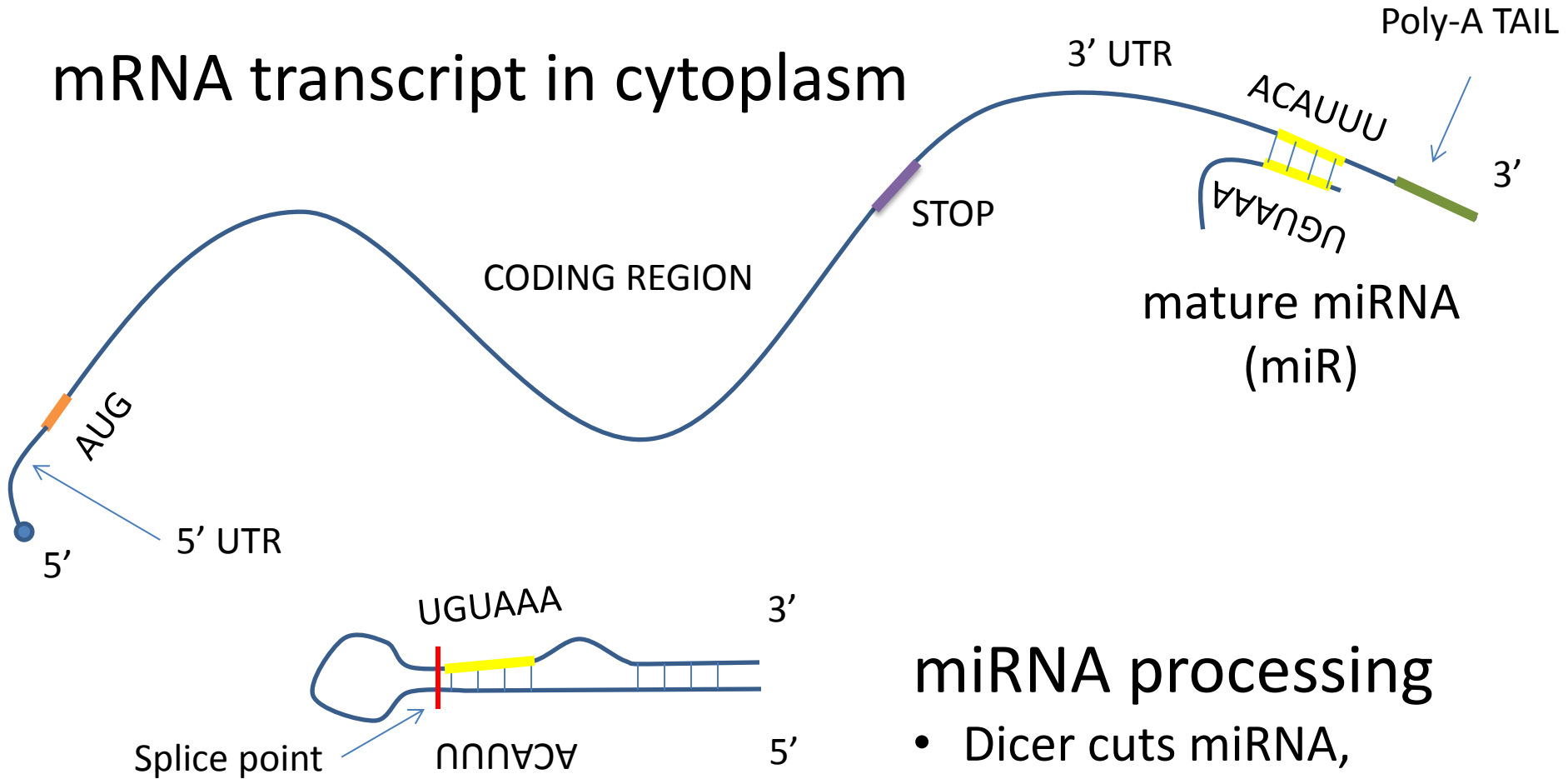
Mentors: Dr. Kevin Chen, Liyang Diao

miRNA

- First discovered in 1993 by Dr. Ambros in *Caenorhabditis elegans*
- Class of small non-coding RNA
- ~22 nt (nucleotides) in length
- Regulates gene expression by binding to mRNA transcript inside cytoplasm
- Acts on 3' untranslated region (3' UTR) of target mRNA

miRNA

mRNA transcript in cytoplasm



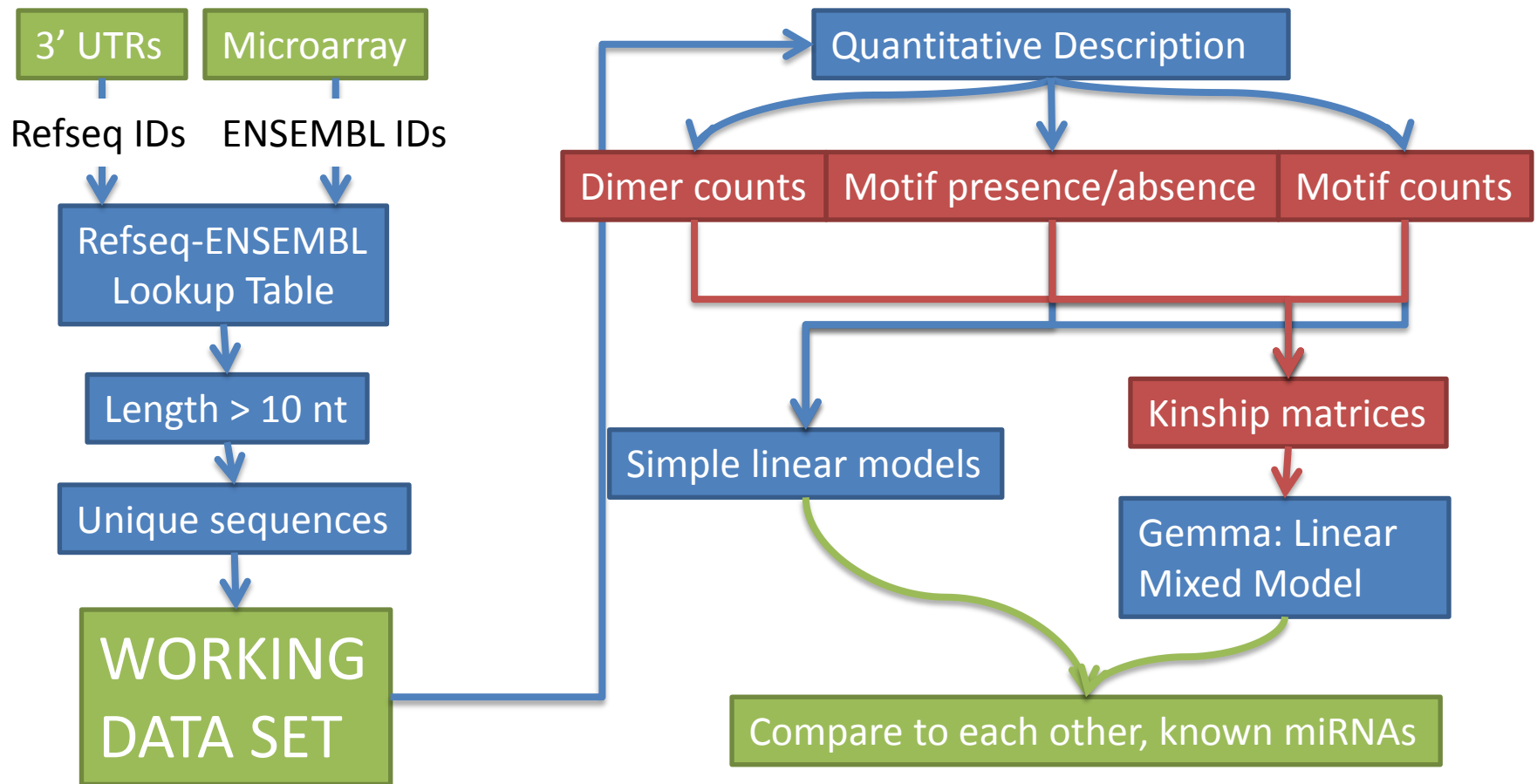
miRNA processing

- Dicer cuts miRNA, activates
 - Disabled in knockout mice

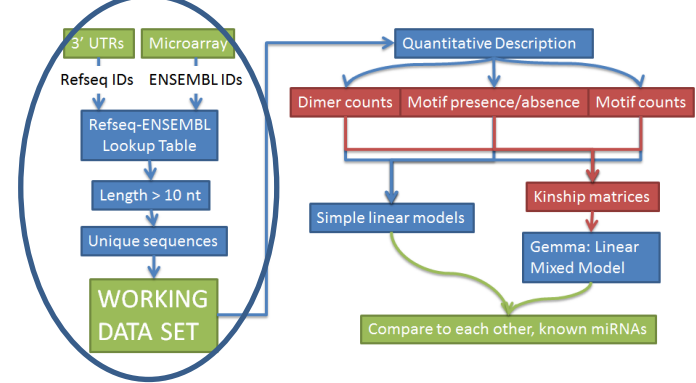
Objective

- Develop a statistical method for discovering new miRNA motifs (6-nt words) based on gene expression arrays
- Desired qualities compared to simple linear model
 - Replicate established results
 - Faster run time
 - Accurately identifies motifs that a simple linear model could not

Procedure overview

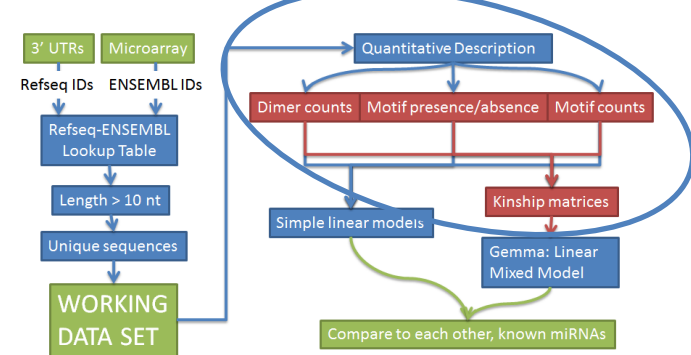


Source data



- Sequences of 3' UTRs from mouse genes
- Microarray data from mouse gene expression
- Criteria for usability
 - Gene in intersection of data sets
 - 3' UTR is at least 10 nt in length
 - 3' UTR is not a substring of another gene's 3' UTR
- Total of 18090 usable genes

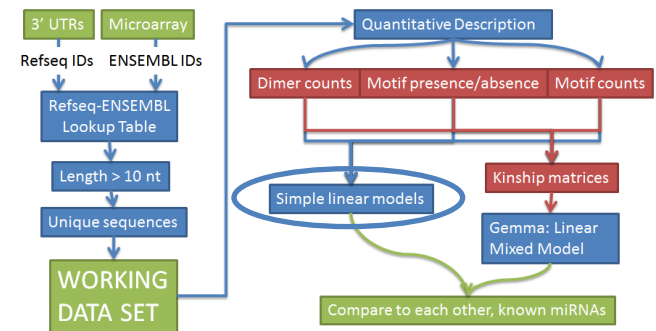
Data preparation



- Quantitative descriptions of source data
 - Dimer counts within each 3' UTR
 - Sixmer counts
 - Presence/absence of each sixmer (bin)
- Kinship matrix
 - Correlate dimer counts
 - Correlate sixmer counts
 - Shared sixmers ($\text{bin}^T * \text{bin}$)

Simple model

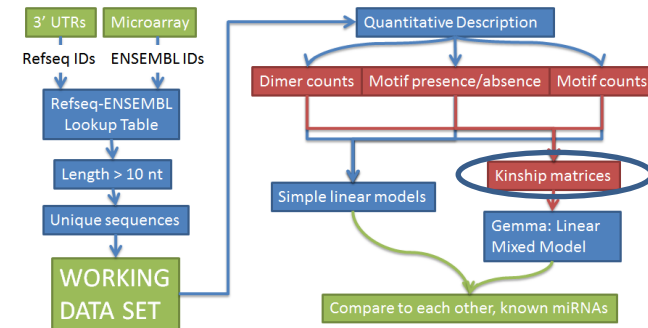
- Simple linear correlation of sixmer presence/absence
 - Pearson correlation:



$$\rho = \frac{E[XY] - E[X]E[Y]}{\sqrt{(E[X^2] - E[X]^2)(E[Y^2] - E[Y]^2)}}$$

$$P = \int_0^{1-\rho^2} t^{\frac{n-4}{2}} (1-t)^{-\frac{1}{2}} dt$$

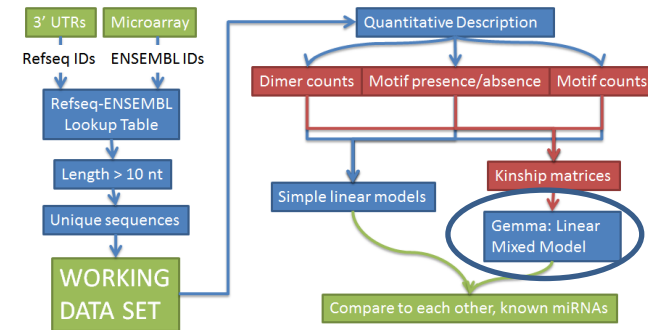
The Kinship Matrix



- Expect more similar 3' UTRs to map to more similar phenotypes
- Correction for relatedness to highlight random effects
- Prefer results to be highly significant ($p < 10^{-5}$) while having high random effect

GEMMA

- Fits linear mixed model of form
- $y = W\alpha + x\beta + u + \varepsilon$
 - $u \sim MVN_n(0, \lambda\tau^{-1}K)$
 - $\varepsilon \sim MVN_n(0, \tau^{-1}I_n)$
 - β , effect size of marker
 - u represents random effects
 - K is the kinship matrix
 - Parameter λ , coefficient of random effect



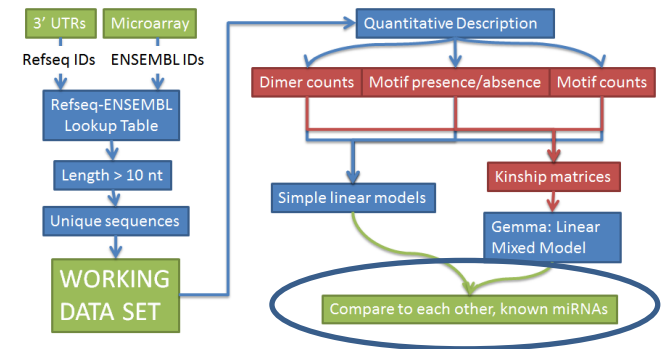
Other linear models

- Simple linear correlation of sixmer counts
- Correction by removing noisy points
 - counts close to 0 should have less effect on expression ratio

Criteria for significant improvement of experimental model

- At least 25% deviation from simple model in top 30 results
- At least 25% increase in number of matches with published miRNA motifs
- Optional: Random effect (λ) is at least 0.1 times the marker effect (β)

Results: Two methods better than control

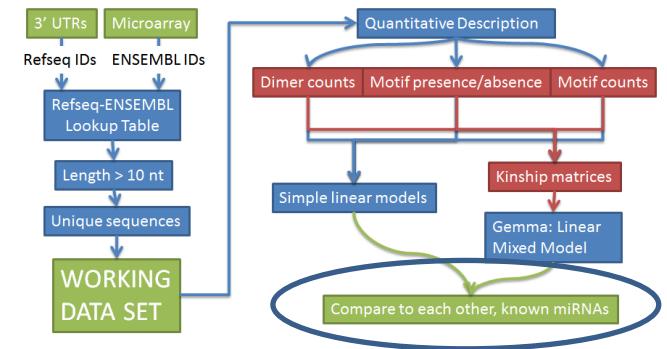


- We have found two statistical methods that have identified motifs that the simple model could not.
- Significant improvement over simple linear model
 - 80% of results returned with the dimer counts matrix but not by the control, had matches in miRBase

Examples of top hits

- UUA AAAA
 - mmu-miR-290b-5p
 - Found on chromosome 7
 - No target prediction available
- UUU AAAA
 - mmu-miR-3094-3p
 - Found on chromosome 4
 - ~ 50 predicted targets at 95%

Methods that did not improve on control



- Shared motifs output nearly identical to simple model output
 - Failed criteria of significant improvement
 - Relative size of random effect $\sim 10^{-2}$
- Threshold counts output: most results did not match other tests
 - However, only 5 of the top 25 matched miRBase – less than control (9/25)
 - Model performance is worse than control

Can we do better?

- Try different measures of kinship
 - k-mer correlation counts with different k
 - Pairwise sequence alignment scores



References

- Lee RC, Feinbaum RL, Ambros V (December 1993). "The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*". *Cell* **75** (5): 843–54.
- Bussemaker, H. J., Li, H., & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature*, 167-171.
- Ben-Moshe, N. B., Avraham, R., Kedmi, M., Zeisel, A., Yitzhaky, A., Yarden, Y., & Domany, E. (2012, August 29). *Context Specific MicroRNA Analysis (CoSMic): Identification of Functional MicroRNAs and their mRNA targets - Supplementary Data*. Retrieved from Nucleic Acids Research: Oxford Journals: <http://nar.oxfordjournals.org/content/suppl/2012/08/29/gks841.DC1/nar-00762-n-2012-File009.pdf>
- Chen, Uing, Huang, Zhouqing, Wang, Liansheng, Wang, Yue, Wu, Feizhen, Meng, Shu, & Wang, Chanqian. "MicroRNA-125a-5p partly regulates the inflammatory response, lipid uptake, and ORP9 expression in oxLDL-stimulated monocyte/macrophages". *Cardiovascular Research*, Oxford Journals. 12 June 2009, Vol 83.1 pg 131-139, at <http://cardiovascres.oxfordjournals.org/content/83/1/131>
- Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis (GEMMA) for association studies. *Nature Genetics*, 44: 821-824, 2012
- Xiang Zhou, Peter Carbonetto and Matthew Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genetics*. in press. <http://arxiv.org/abs/1209.1341>

backup

Results – top 20 most significant

Simple Linear	p_score	Shared sixmers	p_score	Dimer counts	p_score	Correlation of nonzero sixmer counts	p_score	Sixmer counts	p_score
UUAAAA	6.68E-78	UUAAAA	1.53E-74	UUUACA	7.48E-29	AAUAAA	1.95E-31	UGCAAU	5.44E-06
UUUACA	2.38E-74	UUUACA	3.32E-71	UGUAAA	2.63E-27	UUAAAA	8.25E-26	UUAUUA	1.40E-05
UUUAUA	8.62E-74	UUUAUA	1.26E-70	AUCAAA	7.56E-26	UUUAAA	7.39E-25	GUGCAA	1.83E-05
UGUAAA	1.66E-71	UGUAAA	1.78E-68	UUUAUA	1.48E-25	UUUAUU	1.32E-24	CUGUAC	3.10E-05
UUUAAA	2.82E-71	UUUAAA	3.39E-68	AGUUUU	8.97E-25	UAAAAA	2.61E-24	AUCAAA	7.75E-05
UAUAAA	9.27E-70	UAUAAA	1.33E-66	UUAAAA	1.24E-24	UUUUUA	1.49E-23	UUUAUA	8.31E-05
UUUUUA	9.49E-69	UAAAAU	9.55E-66	AAAAGC	1.24E-24	AUUUUU	5.17E-22	CACUGC	8.33E-05
UAAAAU	1.56E-68	UUUUUA	1.27E-65	UGCAAU	5.27E-24	UAUUUU	1.79E-21	UGUAAA	1.06E-04
AUUUUA	5.36E-68	AUUUUA	2.51E-65	UUUUUU	6.35E-24	AAAAAU	2.33E-21	AUUGAU	1.15E-04
UUUUUA	5.71E-68	UUUUUA	5.09E-65	UAUAAA	6.86E-24	AUUUUA	4.58E-21	GCAGUU	1.46E-04
AGUUUU	8.09E-67	AGUUUU	5.44E-64	UUUUAG	1.06E-23	UUUUUA	1.48E-20	UUUACA	1.51E-04
AAAUUA	8.24E-67	AAAUUA	7.63E-64	UUACAU	1.17E-23	AAAACA	1.70E-20	UAAUAU	2.18E-04
UUUUUU	1.21E-65	UUACAU	6.77E-63	AGGAAA	2.45E-23	AUUAAA	2.11E-20	CCGUUG	2.40E-04
UUACAU	1.36E-65	UUUUUU	8.69E-63	AAGAAA	2.72E-23	UUUUUA	2.16E-20	ACGUCC	2.83E-04
AAAAUU	2.70E-65	AAAAUU	1.38E-62	CUUAAA	4.79E-23	ACAAAA	2.90E-20	CAUGAU	2.89E-04
AAAAAU	3.35E-65	AAAAAU	2.81E-62	UAAUAU	1.77E-22	AAAAUA	3.09E-20	GCCAGC	3.32E-04
CUUAAA	1.30E-64	CUUAAA	6.04E-62	UUCCUU	1.81E-22	AAAUAA	5.93E-20	UAGUGU	3.32E-04
UGUUUU	1.74E-64	UAAAAA	1.84E-61	CUUUUU	1.86E-22	UUAUUU	1.43E-19	GCACUA	3.58E-04
UAAAAA	3.47E-64	UGUUUU	2.50E-61	UGUUUU	1.99E-22	UGUUUU	1.57E-19	UUACAU	3.75E-04
UUUUAG	5.11E-63	UUAAAA	3.53E-60	UUUGCA	2.65E-22	AAAAAG	2.08E-19	AUUACA	4.56E-04