Proximity Search Method for Mining Biomedical and Genomic Information

Soma Mbadiwe, Jeremy Dawson, & Don Adjeroh
Computer Science and Electrical Engineering
WestVirginiaUniversity

Presented at the
BigData Network in Biology and Medicine: Modeling Analysis and Challenges
Workshop
IEEE BIBM 2019 Conference
at San Diego, CA, USA

November 18, 2019

The Problem

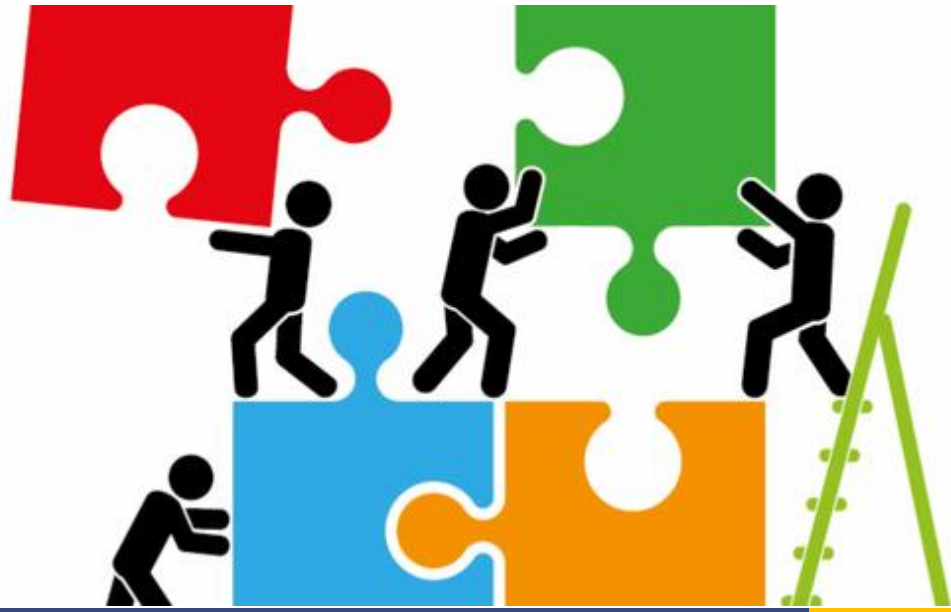
- There are about 7000 known genetic syndromes, with about 2000-3000 of them having some form of facial dysmorphisms (Hart & Hart, 2009).
- Information about genetic syndromes and genes/variants associated with them are mostly in free unstructured text.
  - which is growing exponentially by the day
- How do we effectively identify relevant literature and get these information out without having to read them all?

Proximity Search: Assumptions

- The proximity of the words in a document implies a relationship between the words.
- Authors of documents try to formulate sentences which contain a single idea, or cluster of related ideas within neighboring sentences or organized into paragraphs

Contributions

- A simple but effective pipeline for mining genomic information from literature.
- Chromosomes that are better at discriminating between populations
- Highlighting the effectiveness of proximity matching for quick literature review and establishing relevance



The Task

Identify and select a small panel of SNPs associated with the human face

Motivation for Task

- About 3K genetic syndromes affecting the face
- The face hosts a lot of important functional systems: olfaction, hearing, seeing, etc.
- The face is also important for identification.
- Clear link between facial appearance and genetics.

Approach

- Initial Data Cleaning (Stages 1 & 2)
- Literature Search (Stage 3)
- Proximity Search (Stage 4)
- SNP Ranking (Stage 5)

Dataset Used

- 1000 Genome Project, Phase III dataset
- Contains information on 84.4 million variants (SNPs) from all 24* chromosomes for 2504 subjects
- Subjects from 26 different sub-populations from 5 continents

POPULATIONS AND CONTINENTS REPRESENTED IN SAMPLES FROM 1000 GENOME PROJECT

Code	Ethnicity	Continent	Ethnicity Sample Size	Continent Sample Size
PUR	Puerto Rican	America	104	347
CLM	Colombian	America	94	
PEL	Peruvian	America	85	
MXL	Mexican-American	America	64	
GBR	British	Europe	91	503
FIN	Finnish	Europe	99	
IBS	Spanish	Europe	107	
CEU	CEPH	Europe	99	
TSI	Tuscan	Europe	107	
CHS	Southern Han Chinese	E. Asia	105	504
CDX	Dai Chinese	E. Asia	93	
KHV	Kinh Vietnamese	E. Asia	99	
CHB	Han Chinese	E. Asia	103	
JPT	Japanese	E. Asia	104	
PJL	Punjabi	S. Asia	96	489
BEB	Bengali	S. Asia	86	
STU	Sri Lankan	S. Asia	102	
ITU	Indian	S. Asia	102	
GIH	Gujarati	S. Asia	103	
ACB	African-Caribbean	Africa	96	661
GWD	Gambian	Africa	113	
ESN	Esan	Africa	99	
MSL	Mende	Africa	85	
YRI	Yoruba	Africa	108	
LWK	Luhya	Africa	99	
ASW	African-American SW	Africa	61	

NB: Codes used for the continents in text: AMR for America; EUR for Europe; EAS for East Asia; SAS for South Asia; and AFR for Africa

Approach

- Initial Data Cleaning (Stages 1 & 2)
- Literature Search (Stage 3)
- Proximity Search (Stage 4)
- SNP Ranking (Stage 5)

Initial Cleaning (Stage 1)
– Focus on the SNPs.

- Remove SNPs with **more than one** reference (REF) or alternate (ALT) alleles
  - (e.g., **pos**: 51714 **rsID**: rs52826 **ref**: AAGTT **alt**: A)
- Remove SNPs that have the **same REF and ALT alleles**
  - (e,g., **pos**: 15903 **rsID**: rs557514207 **ref**: G **alt**: G)

Initial Cleaning (Stage 2)
– Focus on the Individuals.

- Remove SNPs that all subjects have the same REF/ALT values
- For each SNP, extract their position/loci number, rsID, reference (REF) allele, alternate (ALT) allele and allele information of all 2504 subjects

Approach

- Initial Data Cleaning (Stages 1 & 2)
- Literature Search (Stage 3)
- Proximity Search (Stage 4)
- SNP Ranking (Stage 5)

Literature Search

- Search focused on PubMed
  - Largest repository of medical and genomic publications.
- 89,752 PubMed queries
  - 11.216 genes * 8 face regions
- Query Format:

  gene_name ***AND*** *(face_region* ***OR*** *alt_term_1* ***OR*** *alt_term_2* ***OR*** *…* ***OR*** *alt_term_n)*

  e.g: PAX3 AND (cheek OR buccal OR nasolabial OR cheekbone)

Face Regions and Alternative Terms Used

Face region	Alternative terms
eye	iris, pupil, sclera, eyebrow, eyelid, orbit, cornea, lens, epicanthic, eyelash, orbicularis oculi, ocular, visual
nose	nostril, ala nasi, nasalis, nasal, glabella, maxilla
mouth	rima oris, lip, philtrum, orbicularis oris, oral, dental
ear	earlobe, pinna, auricle, auricula, tragus, antitragus, auditory
cheek	nasolabial, buccal, cheekbone
chin	mentolabial sulcus, mandible, jaw, mental region, mandible
forehead	temple, frontalis, brow ridge, supraorbital ridge, superciliary arch
face shape	facial shape, facial morphology, cranium, cranio

Approach

- Initial Data Cleaning (Stages 1 & 2)
- Literature Search (Stage 3)
- Proximity Search (Stage 4)
- SNP Ranking (Stage 5)

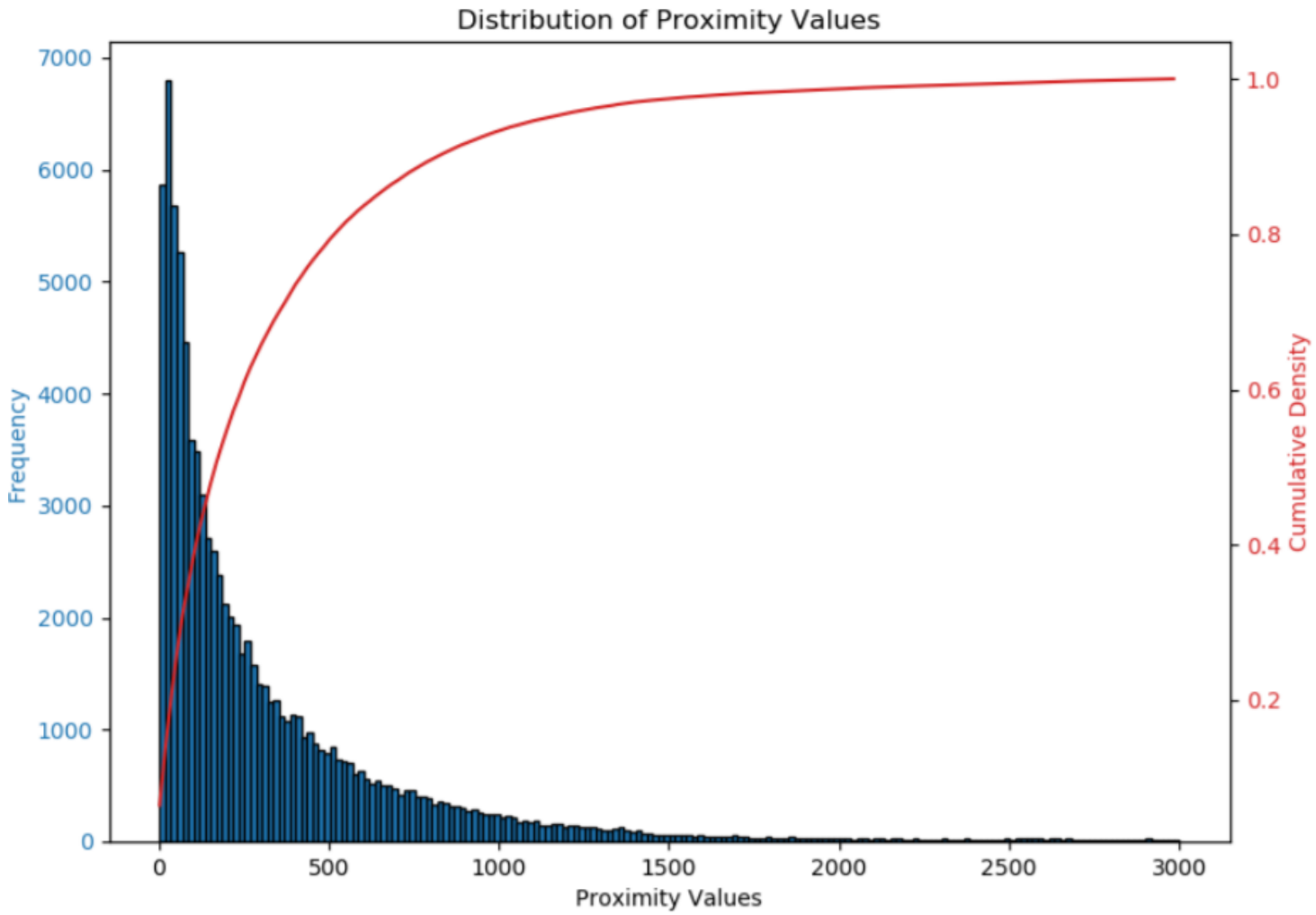
Proximity Search: Approach

- Take a text file containing title and abstract, or title, abstract and text body.
- Construct a suffix tree with the file content. [$\Theta(n)$-time]
- Get all the genes and face regions that pointed to the file.
- For each gene - face region pair, we build a list containing the region and its alternative terms
  - e.g. **Gene**: PAX3; **List**: [cheek, buccal, nasolabial, cheekbone]

Proximity Search: Approach

- Search the suffix for all occurrences of the gene and each item in the list.
  - **Output**: G: list of indexes where the gene occurs in the document; F: list of indexes where the face regions occurs in the document;
- Use a modified binary search algorithm to find an occurrence of the gene, G_i, and an occurrence of any of the items in the face regions list, F_j
  - such that $\eta = |G_i - F_j|$ is minimized
- Return η as the proximity value for the gene-face region pair.
  - $\eta \in \mathbb{Z}^+$.
  - If the gene or any of the face region terms is not found in th tree, we set η to -1 to mark it as invalid.

Proximity Values: Summary



Count: 94508

Max: 76389

Min: 3

Mean: 726.9

Median: 197

Searched: 36,336 full papers and 103,190 title-and abstract-only

Search: Highlights

- 4,646 (41%) genes queried returned one or more hits.
- Chromosome 2 got the most hits in absolute terms.
- Chromosome 21* has the highest hit ratio (0.7143):
  - hit ratio $= \frac{\#\ genes\ with\ hits}{\#\ genes\ queried}$
- *Hits* here means those with valid proximity values

SUMMARY OF RESULTS FOR DATASET AND LITERATURE SEARCH

Chr	# variants (in 1KGP)	# SNPs (in 1KGP)	# genes queried	# genes with hits	# SNPs in files
1	6,468,094	6,215,039	1,174	317	421
2	7,081,600	6,807,712	735	442	368
3	5,832,276	5,602,392	605	408	248
4	5,732,585	5,499,209	445	146	198
5	5,265,763	5,054,711	511	123	181
6	5,024,119	4,816,043	600	385	407
7	4,716,715	4,532,453	499	273	350
8	4,597,105	4,433,737	425	272	218
9	3,560,687	3,426,741	410	155	151
10	3,992,219	3,836,556	415	209	276
11	4,045,628	3,890,940	695	359	281
12	3,868,428	3,709,639	616	186	210
13	2,857,916	2,736,546	193	100	107
14	2,655,067	2,547,628	344	82	81
15	2,424,689	2,328,197	331	80	203
16	2,697,949	2,606,624	464	195	232
17	2,329,288	2,234,278	708	137	262
18	2,267,185	2,178,378	169	38	63
19	1,832,506	1,758,136	699	258	168
20	1,812,841	1,744,900	305	92	125
21	1,105,538	1,058,339	119	85	112
22	1,103,547	1,059,517	281	98	147
X	3,468,093	3,238,982	467	199	97
Y	62,042	60,505	9	7	0
Total	84,801,880	81,377,202	11,219	4,646	4,906

Column *"# genes queried"* shows the number of genes that were used in the literature search on PubMed. Column *"# genes with hits"* shows the number of genes that appeared with at least a face region in at least one publication. Column *"# SNPs in files"* shows the number of SNPs found by searching for rsIDs in the publications that has a valid proximity value. *"1KGP"* refers to the 1000-Genome Project dataset.

Approach

- Initial Data Cleaning (Stages 1 & 2)
- Literature Search (Stage 3)
- Proximity Search (Stage 4)
- SNP Ranking (Stage 5)

Data Pruning Stages: Summary

9.4M SNPs across chromosomes

PRUNING SUMMARY USING DATA FROM THE 1000 GENOME PROJECT

Chr	Stage 1	Stage 2	Stage 3	Stage 4
1	6,196,151	6,189,941	1,636,512	518,073
2	6,786,300	6,779,891	1,490,802	975,857
3	5,584,397	5,579,072	1,276,695	947,717
4	5,480,936	5,475,541	1,039,230	391,508
5	5,037,955	5,034,476	1,126,224	338,140
6	4,800,101	4,796,486	1,141,240	794,531
7	4,517,734	4,514,054	1,207,843	763,096
8	4,417,367	4,413,082	944,929	670,212
9	3,414,848	3,411,936	909,246	349,289
10	3,823,786	3,819,395	1,015,610	592,330
11	3,877,543	3,872,914	1,049,921	580,787
12	3,697,856	3,694,543	928,208	320,699
13	2,727,881	2,724,583	547,922	319,911
14	2,539,149	2,535,345	590,453	186,600
15	2,320,474	2,318,207	760,828	233,727
16	2,596,072	2,593,765	703,361	321,293
17	2,227,079	2,174,862	761,605	158,310
18	2,171,378	2,169,126	457,391	94,179
19	1,751,878	1,750,709	532,527	193,786
20	1,739,315	1,737,721	411,308	116,340
21	1,054,447	1,051,751	303,088	250,328
22	1,055,454	1,054,440	410,326	176,116
X	1,391,537	1,391,235	271,479	135,369
Y	6,389	6,255	539	496
Total	79,216,027	79,089,330	19,517,287	9,428,694

Stage 1: data after removing all variants that are not SNPs, or with no REF, or with more than one ALT, or where the REF allele is also in the ALT alleles. **Stage 2**: data after those where all individuals have the same allele were removed. **Stage 3**: data after SNPs not within a gene were removed. **Stage 4**: data after those SNPs that are not part of the genes found from proximity search were removed.

#5: SNP Ranking

- For each gene that meets a given proximity threshold:
  - Compute and rank the SNPs by their F_{ST} values
  - Compute and rank the SNPs by their I_n values
  - Take the top 100 from each ranking
  - Merge the two sets of SNPs
- Combine all the selected SNPs from each gene and rank again by F_{ST} and by I_n
- Take the top 200 from each ranking and merge into the final set.

Informativeness (Rosenberg et al. 2003)

Best understood as the expected log of the likelihood that an allele is assigned to one of the populations compared with a hypothetical "average" population whose allele frequencies equal the mean allele frequency across all the populations

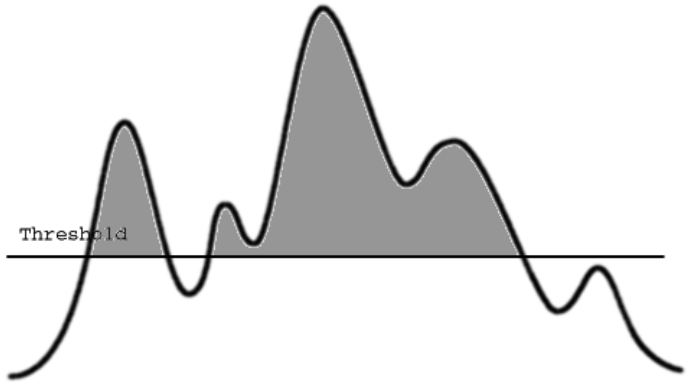
$$I_n = \sum_{j=1}^{N} \left(-p_j \log p_j + \frac{1}{K} \sum_{i=1}^{K} p_{ij} \log p_{ij} \right)$$

p_{ij} is frequency of allele j in population i.

p_j is the average frequency of allele j over the K populations

Proximity Thresholds considered

- $\eta = 50$
  - Roughly a sentence.
- $\eta = 500$
  - Roughly a couple of sentences.
- $\eta = 1500$
  - Roughly a paragraph.
- $\eta \to \infty$
  - No threshold. Use all.



PER-CONTINENT CLASSIFICATION ACCURACY USING DIFFERENT PROXIMITY THRESHOLDS (η)

	th=None		th=1500		th=500		th=50		SNPs in file*	
Chr	# SNPs	Accuracy	# SNPs	Accuracy	# SNPs	Accuracy	# SNPs	Accuracy	# SNPs	Accuracy
1	258	92.21 ± 0.95	247	91.25 ± 1.10	244	90.02 ± 1.34	240	87.94 ± 1.17	239	91.29 ± 1.04
2	245	88.14 ± 1.15	244	87.42 ± 1.67	244	86.58 ± 0.62	245	88.02 ± 1.97	303	**93.33 ± 0.74**
3	248	**94.01 ± 1.30**	249	**93.25 ± 0.92**	236	92.01 ± 1.09	236	**92.05 ± 1.29**	209	91.01 ± 1.52
4	253	89.50 ± 2.28	253	89.86 ± 0.51	260	89.89 ± 0.93	239	88.66 ± 1.56	113	88.78 ± 0.75
5	238	86.62 ± 0.91	238	87.50 ± 0.59	233	86.54 ± 0.93	248	85.82 ± 0.35	90	82.55 ± 1.81
6	243	87.26 ± 0.67	243	86.10 ± 1.65	256	84.55 ± 1.23	260	85.34 ± 0.94	353	92.33 ± 1.22
7	221	84.70 ± 0.33	224	85.58 ± 1.15	225	84.87 ± 1.54	234	88.02 ± 1.05	290	92.57 ± 0.81
8	239	85.46 ± 0.46	240	85.22 ± 1.64	241	85.06 ± 1.42	224	83.86 ± 0.57	192	89.50 ± 0.72
9	231	88.50 ± 0.89	236	89.38 ± 0.94	242	88.77 ± 1.67	235	88.54 ± 0.85	95	83.63 ± 0.89
10	239	84.75 ± 1.31	243	87.58 ± 1.16	230	85.86 ± 1.73	242	90.26 ± 1.26	222	88.98 ± 1.57
11	247	90.38 ± 0.97	231	92.57 ± 1.51	229	92.09 ± 0.77	253	90.69 ± 1.12	222	91.37 ± 0.63
12	233	92.05 ± 0.76	243	92.25 ± 1.16	241	92.18 ± 1.61	215	89.10 ± 1.86	114	87.38 ± 1.65
13	228	84.82 ± 1.01	227	85.19 ± 1.18	227	84.42 ± 0.52	222	82.75 ± 1.46	75	81.23 ± 1.04
14	224	89.10 ± 0.87	224	88.70 ± 1.35	226	88.85 ± 2.72	242	88.70 ± 0.85	46	78.87 ± 1.32
15	255	92.81 ± 1.18	255	92.77 ± 0.93	248	92.05 ± 0.96	224	88.22 ± 0.52	88	89.02 ± 1.15
16	230	93.57 ± 1.02	230	93.21 ± 0.75	246	**93.69 ± 0.74**	228	91.14 ± 0.86	151	90.50 ± 1.31
17	235	75.80 ± 1.49	227	77.12 ± 1.31	228	80.39 ± 0.60	227	86.06 ± 0.97	115	83.50 ± 2.67
18	208	73.69 ± 2.23	208	73.84 ± 1.91	209	77.48 ± 2.12	211	81.75 ± 1.06	31	70.49 ± 0.87
19	236	87.18 ± 1.16	234	86.74 ± 1.31	231	86.30 ± 1.24	253	89.66 ± 1.28	103	85.70 ± 1.01
20	232	85.94 ± 1.50	232	86.46 ± 1.34	231	86.74 ± 0.86	239	86.26 ± 1.28	58	77.23 ± 1.04
21	245	86.22 ± 0.98	245	85.82 ± 1.02	242	84.74 ± 0.87	224	84.23 ± 1.20	89	81.15 ± 1.34
22	221	88.34 ± 0.87	239	89.34 ± 0.57	235	90.10 ± 0.45	234	88.50 ± 1.07	85	82.43 ± 1.08
X	220	56.47 ± 2.14	217	58.22 ± 1.35	220	58.15 ± 1.77	223	59.06 ± 1.74	62	56.31 ± 1.80
Y	213	35.92 ± 1.99	213	36.25 ± 1.08	208	34.79 ± 1.83	112	29.04 ± 0.75	0	-
	5642	83.89 ± 1.19	5642	84.23 ± 1.17	5632	84.01 ± 1.23	5510	83.90 ± 1.13	3345	84.75 ± 1.22

* These are a subset of SNPs found in the publications with valid proximity values. Interestingly, all the publications in which we found a SNP have proximity values $\eta \leq 500$. Threshold values, th, specify the upper bound for proximitiy value. E.g. $th = 500 \implies 0 < \eta \leq 500$.

GENE COUNT AS PROXIMITY VALUE THRESHOLD CHANGES

Chr	th=None	th=3000	th=2500	th=2000	th=1500	th=1000	th=500	th=300	th=200	th=100	th=50	th=30
1	317	277	273	269	263	243	215	175	149	105	66	39
2	442	426	425	424	420	409	379	341	305	256	191	152
3	408	394	393	390	386	379	348	312	284	226	173	127
4	146	138	138	137	137	127	121	104	84	55	42	30
5	123	119	118	118	117	112	99	88	81	67	36	26
6	385	372	371	368	366	359	328	295	270	220	165	138
7	273	258	256	255	253	248	221	196	172	145	119	90
8	272	259	259	259	259	247	229	195	181	147	121	103
9	155	153	153	152	151	148	136	128	111	92	70	50
10	209	195	191	188	187	182	166	147	135	107	85	69
11	359	343	342	341	337	322	294	257	236	196	150	100
12	186	166	162	156	156	150	124	106	90	65	43	25
13	100	93	93	90	90	86	78	62	52	42	28	16
14	82	75	74	71	71	70	65	53	47	35	25	17
15	80	79	79	79	79	78	62	50	45	35	28	22
16	195	190	186	186	185	184	170	156	141	115	94	72
17	137	133	132	132	132	132	111	101	85	65	29	15
18	38	36	36	36	36	35	30	24	22	20	7	2
19	258	249	248	246	242	236	216	203	180	156	122	95
20	92	85	85	84	83	81	79	67	61	46	28	17
21	85	85	83	83	82	81	75	69	67	65	53	41
22	98	90	90	89	87	85	79	72	65	47	39	26
X	199	185	183	181	176	174	149	126	110	77	56	37
Y	7	7	7	7	7	6	5	5	3	2	2	0
Total	4,646	4,407	4,377	4,341	4,302	4,174	3,779	3,332	2,976	2,386	1,772	1,309
Rel Diff		239	30	36	39	128	395	447	356	590	614	463
Abs Diff		239	269	305	344	472	867	1,314	1,670	2,260	2,874	3,337

Rel Diff show the difference between a gene count and the one immediately before it. *Abs Diff* show the difference between a SNP count and the count at th=None. Threshold values, th, specify the upper bound for proximitiy value. E.g. $th = 500 \implies 0 < \eta \leq 500$.

A few take-aways

- Chromosomes 17, 18, X and Y consistently performed worse than the others
- Chromosomes 3 and 16 were among the best performing
- Weak correlation between classification accuracy and number of SNPs use used.
- Sometimes performance improves when threshold is reduced.

Limitations

- Database of published work continues to grow.
  - The new data can affect selection of candidate genes.
- We considered only publications freely available on PubMed / PMC.
  - Title and abstract only for non-free publications
- Selection of literature has limited context-awareness.
  - Possibility of false positives.

Future Work / Final Thoughts

- Approaches to implementing a more context-sensitive approach to understanding literature.
  - This work establishes a baseline to measure effectiveness of more sophisticated approaches
- How can we use the SNPs we've found to improve our understanding of genetic diseases?

Thank You!

Questions?