I023

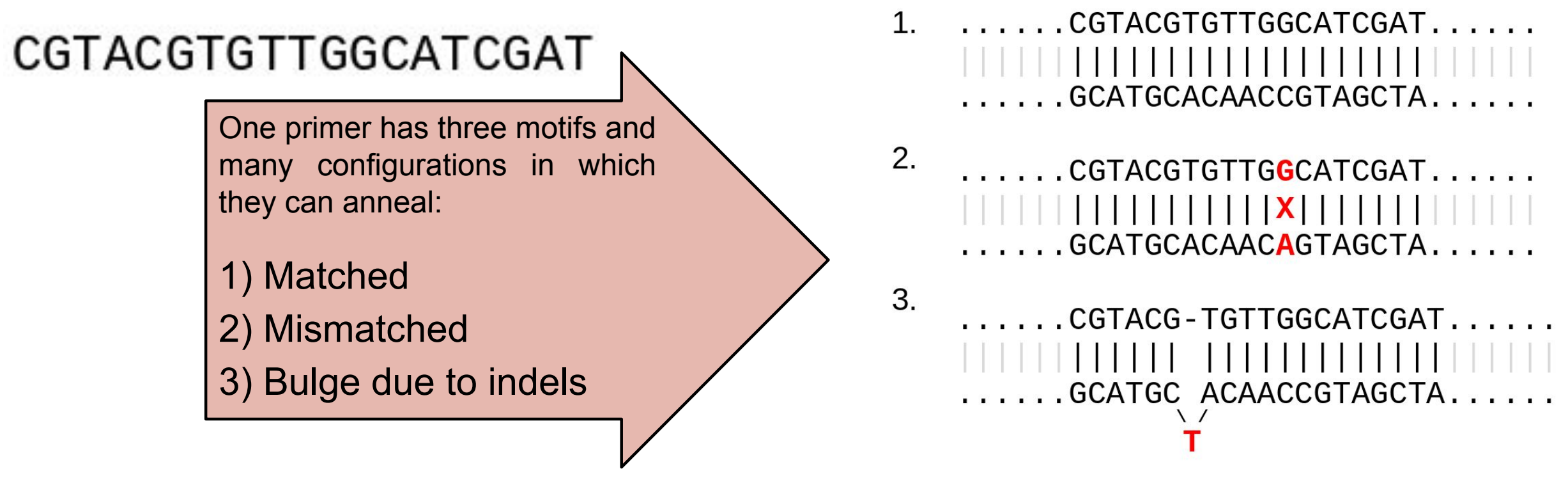
Genomic Database for Assessing Specificity of Primers with Mismatches and Single-Base Bulges

Zachary Dwight, MS, MBA; **Carl Wittwer**, MD, PhD

Department of Pathology, University of Utah, Salt Lake City, UT

Introduction

Good primer design is critical for robust and reproducible PCR. Many factors are important in primer design, leading to a cumbersome number of parameters to input and evaluate. Furthermore, primer specificity is crucial for creating assays that amplify the intended target efficiently while avoiding non-specific products. In recent years, computational techniques have been optimized to quickly search the entire genome for potential primer sites. Unintended perfect matches as well as single base mismatches are typically considered. However, other structures such as single-base bulges are seldom considered. On average, single base bulges are less than half as destabilizing as single base mismatches. Recent studies have emphasized the importance of 3' specificity of a primer rather than considering the entire primer sequence. A genomic database of 14 bp fragments accompanied by positional information was developed and optimized for a straightforward and effective primer search process that includes exact, mismatch, and single-base bulges.



CGTACGTGTTGGCATCGAT

One primer has three motifs and many configurations in which they can anneal:

1) Matched
2) Mismatched
3) Bulge due to indels

```
1.   ......CGTACGTGTTGGCATCGAT......
     |||||||||||||||||||||||||||||||
     ......GCATGCACAACCGTAGCTA......

2.   ......CGTACGTGTTGGCATCGAT......
     |||||||||||||||X|||||||||||||||
     ......GCATGCACAACAGTAGCTA......

3.   ......CGTACG-TGTTGGCATCGAT......
     |||||||||| ||||||||||||||||||||
     ......GCATGC ACAACCGTAGCTA......
                 \/
                 T
```

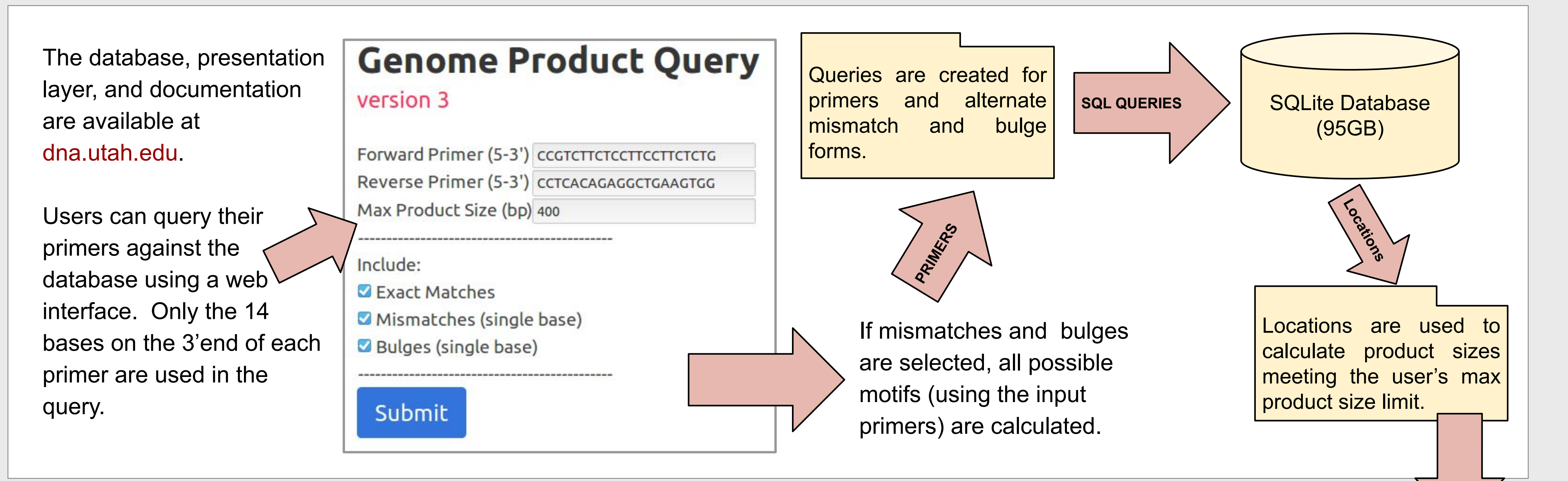
Materials and Methods

The customized genomic database includes all 14 bp fragments (14*mers*) existing in the human genome (GRCh38 reference). All sites were identified, stored and counted via Ruby (https://www.ruby-lang.org) and stored in SQLite 3.2 (https://www.sqlite.org). The final indexed database (with SQL scripts) can retrieve locations, total occurrences and the containing chromosomes. An exact match search was built and additional Ruby scripts developed to iterate possible permutations of mismatches and single-base bulges. Structures that were excluded from results included single base bulges on the 3'-end, 5'-end, and the 3' penultimate position, and 3'-end mismatches. A test set of 170 primer pairs was assessed with this tool as well as external public and web accessible software for benchmarking purposes. Metrics such as genomic site matches and query time required (seconds) were compared.

Fig. 2 -- Contingency Table - Returned # of Potential Products

Total Products	Succesful PCR	Poor PCR	Total	% Succesful
> 15	57	25	82	69.51%
<= 15	79	9	88	89.77%
			170	p < 0.001

Contingency tables and chi-squared statistics were calculated to investigate the association between in-silico genome searches and PCR success. Assays were identified as successful if the intended target was amplified while also avoiding non-specific amplification (confirmed via gel). High-resolution melting curves and thermodynamic predictions (uMelt and Tm Tool) were also used to confirm correct products.



The database, presentation layer, and documentation are available at dna.utah.edu.

Users can query their primers against the database using a web interface. Only the 14 bases on the 3'end of each primer are used in the query.

Genome Product Query
version 3

Forward Primer (5-3') CCGTCTTCTCCTTCCTTCTCTG
Reverse Primer (5-3') CCTCACAGAGGCTGAAGTGG
Max Product Size (bp) 400

Include:
☑ Exact Matches
☑ Mismatches (single base)
☑ Bulges (single base)

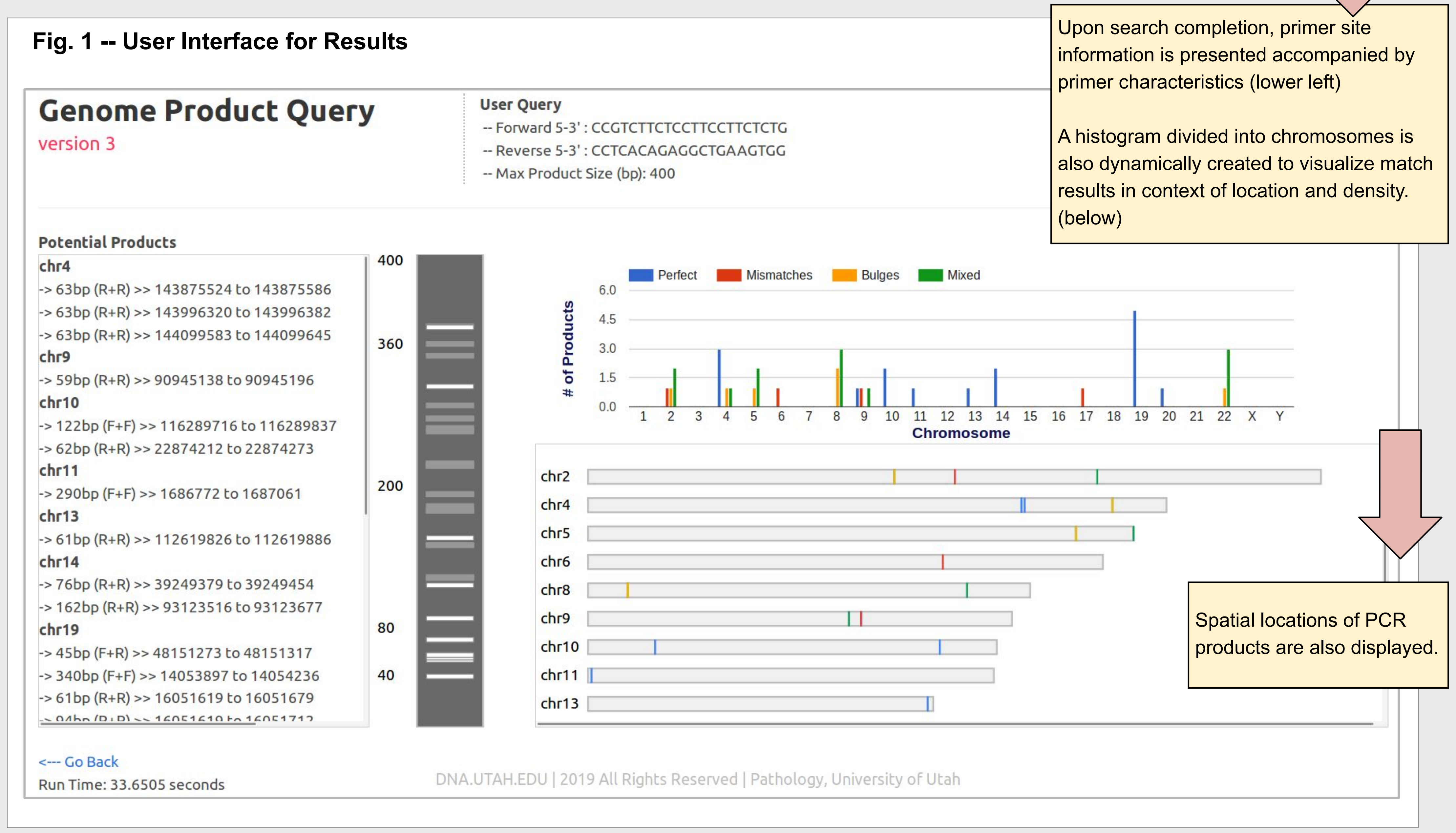
If mismatches and bulges are selected, all possible motifs (using the input primers) are calculated.

Queries are created for primers and alternate mismatch and bulge forms.

SQLite Database (95GB)

Locations are used to calculate product sizes meeting the user's max product size limit.

Fig. 1 -- User Interface for Results



Upon search completion, primer site information is presented accompanied by primer characteristics (lower left)

A histogram divided into chromosomes is also dynamically created to visualize match results in context of location and density. (below)

Genome Product Query
version 3

User Query
-- Forward 5-3' : CCGTCTTCTCCTTCCTTCTCTG
-- Reverse 5-3' : CCTCACAGAGGCTGAAGTGG
-- Max Product Size (bp): 400

Potential Products
chr4
-> 63bp (R+R) >> 143875524 to 143875586
-> 63bp (R+R) >> 143996320 to 143996382
-> 63bp (R+R) >> 144099583 to 144099645
chr9
-> 59bp (R+R) >> 90945138 to 90945196
chr10
-> 122bp (F+F) >> 116289716 to 116289837
-> 62bp (R+R) >> 22874212 to 22874273
chr11
-> 290bp (F+F) >> 1686772 to 1687061
chr13
-> 61bp (R+R) >> 112619826 to 112619886
chr14
-> 76bp (R+R) >> 39249379 to 39249454
-> 162bp (R+R) >> 93123516 to 93123677
chr19
-> 45bp (F+R) >> 48151273 to 48151317
-> 340bp (F+F) >> 14053897 to 14054236
-> 61bp (R+R) >> 16051619 to 16051679
-> 94bp (R+R) >> 16051619 to 16051712

Spatial locations of PCR products are also displayed.

<--- Go Back
Run Time: 33.6505 seconds

DNA.UTAH.EDU | 2019 All Rights Reserved | Pathology, University of Utah

All our software and prototypes can be found at DNA.UTAH.EDU and questions can be sent to zach.dwight@path.utah.edu.

Figure 3 -- Pseudocode for Database Development

All unique 14bp sequences existing in the human genome (GRCh38 reference) were identified, stored and counted via Ruby and SQLite. The process below converted all chromosome (FASTA) files from raw text to a single SQLite database file with a size of 95GB.

```
For each chromosome in genome:
    For each index in chromosome:
        If 14bp sequence not in hash table:
            Store 14bp sequence w/ index location
        Else:
            Add index to existing 14bp sequence locations
    Write hash table to file (.txt)

For each chromosome file (.txt):
    Create new database table for chromosome
    For each row in file:
        Insert row [chromosome, sequence, locations] into database table

For each chromosome table in database:
    Assign 14bp sequence as primary key
    Create table index on primary key
```

To quickly sort and construct genomic data in hash tables, an abundance of memory is required. During the build, as much as 70% of the total RAM (80GB) was utilized simultaneously.

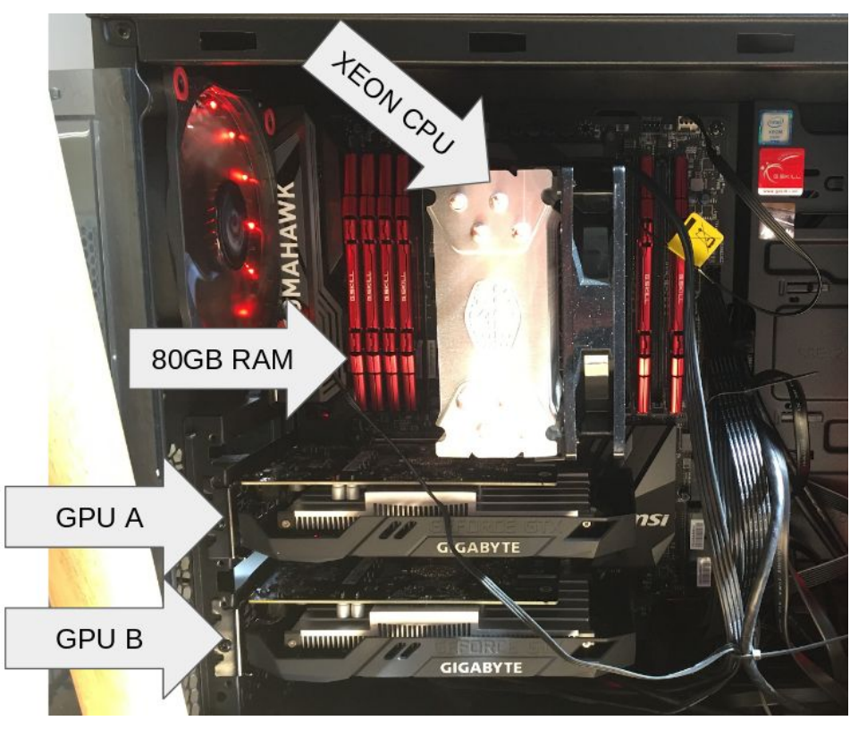


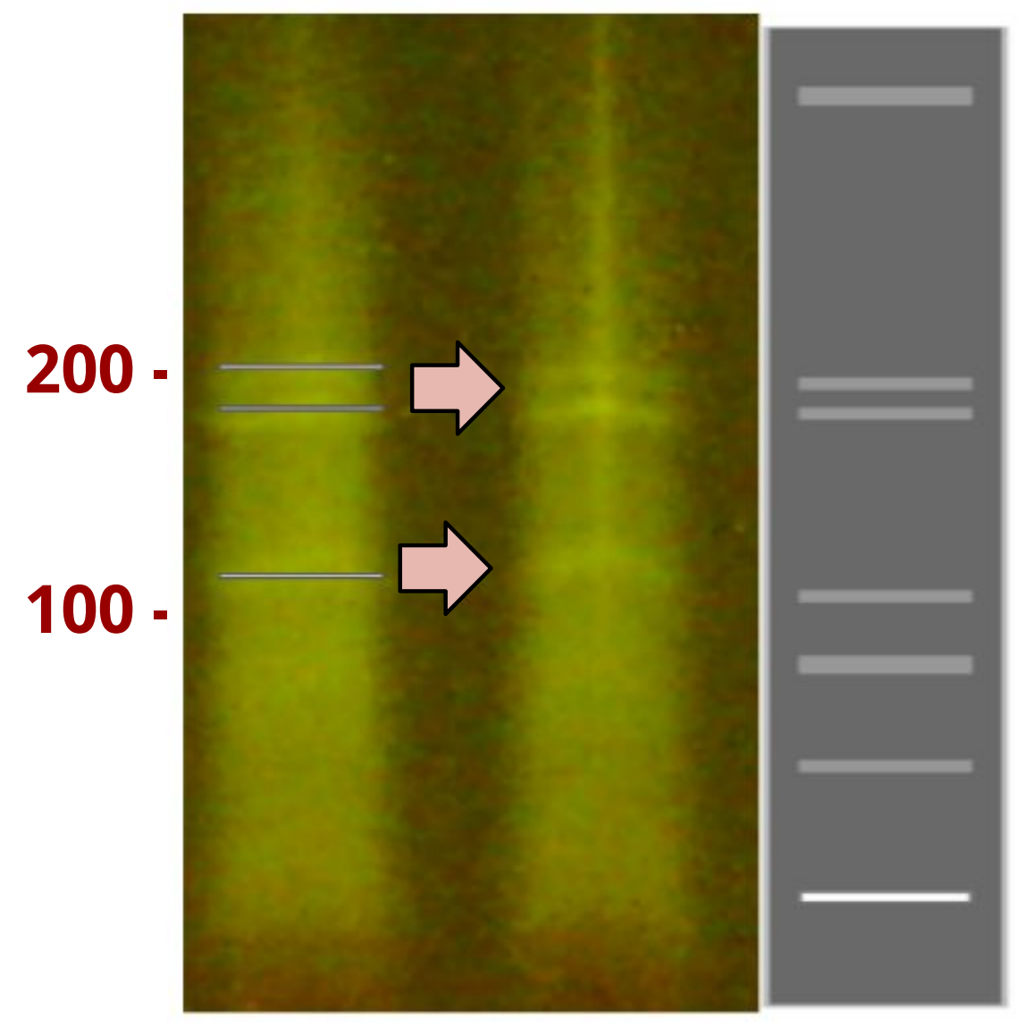
Figure 4 -- Dynamic *In Silico* Gel

Upon search completion, a simulated gel is created and displayed to the user. Products that are perfectly matched to the queried primers are displayed as the brightest bands (white) and alternate products are displayed in different shades of grey. The simulated gel can aid in identifying products that may include bulges and mismatches (below).

Small Amplicon (rs#3128598)
- Product size: 35 bp
  F: GACCTGGCACCACTGC
  R: GGAGTCAGGGCGGAGG

The *in silico* search and simulated gel revealed potential products (grey bands) that match many of the bands on the actual gel and most likely more thermodynamically stable.

The intended 35 bp target appears on the *in silico* gel (white) but does not appear as a distinct band experimentally.



Searching the Genome with GPUs

The growing popularity of graphics processing units (GPUs) have made it easier than ever to perform massive amounts of simple tasks in parallel, drastically reducing computational time.

Fig. 5 -- Comparison of assays: specificity and computational time (sec)

# Potential Sites	CPU	Multi-thread	GPU
~185k	31	20.7	0.6
~ 800k	97.9	59.6	9.4
~ 8 million	233.4	81.7	11.3

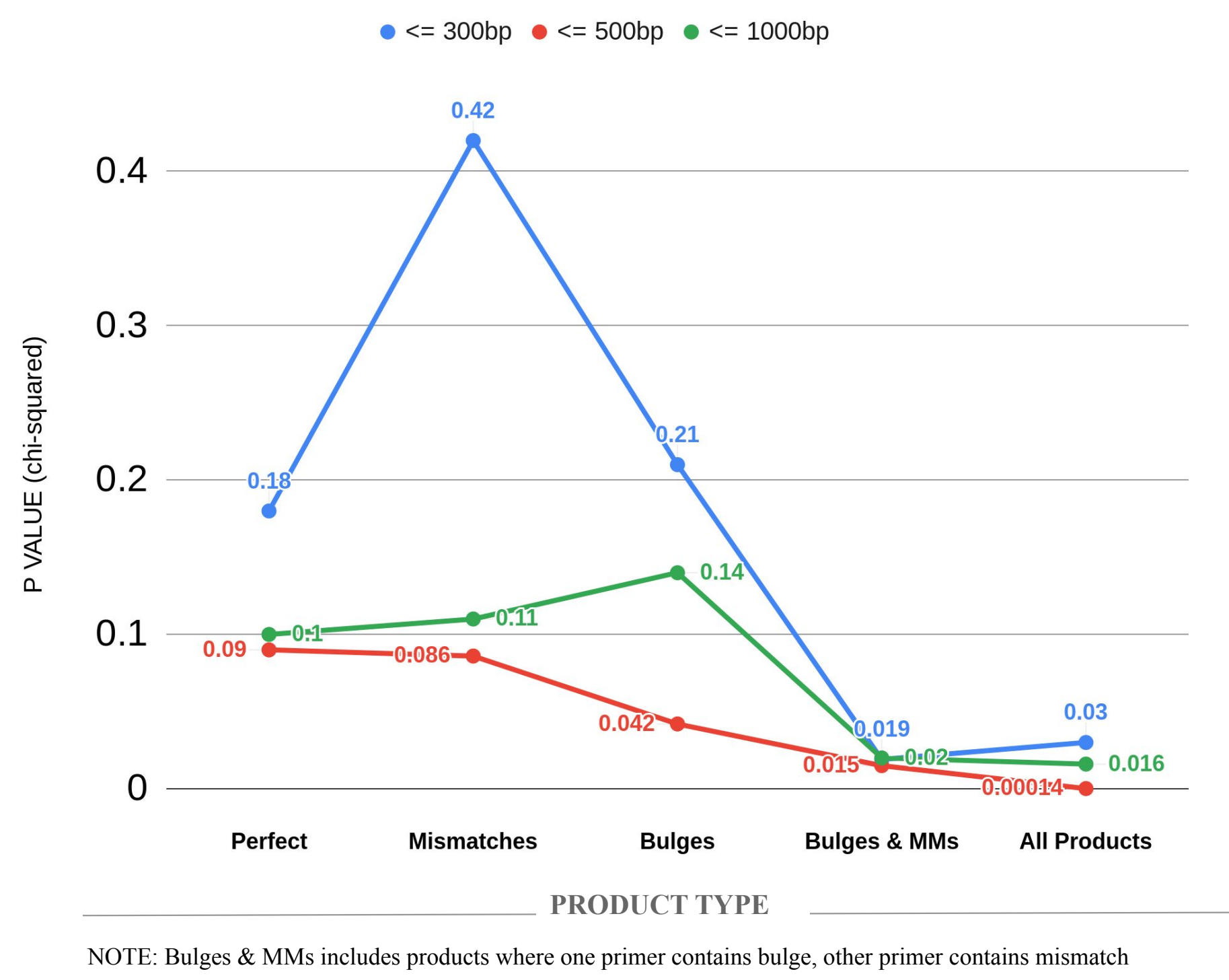
The performance improvement with GPUs is primarily seen when a massive amount of potential products is observed (A/T rich, repeated regions, etc). Future work includes adding multiple bulges and multiple mismatches which will require an extensive number of comparisons and will rely heavily on GPUs.

ADDITIONAL SPECS | **CPU**: Xeon E5-2603V3 LGA2011-3 (6 cores),
RAM (80GB): G.SKILL Ripjaws V Series , **Motherboard:** MSI X99A Tomahawk
GPU: NVidia Geforce GTX 1650 OC (2 cards)

Results

Data from 170 small amplicon assays were assessed for success and failure and compared to genomic search results. Limiting genomic searches to one bulge or one mismatch per primer yielded a chi-square result of **p<0.0002** when assessing total alternate products found (<500bp). On average, the fastest search available with this database (via web server) is an exact match for a primer pair (~0.01s) where mismatches and single-base bulges are excluded. When mismatches are included, the amount of time increases to ~4.5s. WIth all options included, the time via web server is much slower at ~15s. Utilizing GPUs can greatly increase speed given an abundance of sites that are compared.

Fig. 6 -- Comparison of Product Size Limits on Genomic Search Results



NOTE: Bulges & MMs includes products where one primer contains bulge, other primer contains mismatch

The inclusion of single-base bulges in addition to mismatches improved p values. The p values (Fig. 6) describe the association between potentials products returned in a computational search and PCR success in a set of 170 small amplicons.

Conclusion

This database, via web server or GPU version, provides a thorough search for target locations and possible mispriming sites involving many arrangements of mismatches and single-base bulges. Future testing includes identification of statistically significant parameters to better predict primer success after genomic search for unintended binding.

LINK TO APP