Supplementary Figure Legends

Supplementary 1 Recapitulation of a published CRC study

As a proof-of-concept case study, GeMSTONE (1) recapitulated every step in the original Colorectal-Cancer prioritization workflow[1], (2) rescuing 26 out of 28 candidate variants from the ~30,000 variants in the raw whole exome sequencing dataset, and (3) hitting all hereditary CRC and CRC GWAS variants.

Supplementary 2 ALS study workflow

Only simple filters were applied for the ALS case study—specifically, variants were required to be of high quality with (1) genotype quality ≥40, (2) putative damaging effect (frameshift, in-frame indel, non-synonymous, and essential splice site variant) on protein coding transcript, (3) absent from control cohort and (4) rare in the general population defined by an allele frequency ≤0.05%. Two different co-segregation analyses were performed (Dominant and Recessive), in line with ALS inheritance studies[2]. Recurrence filters and family constraints were applied, requiring variant occurrence in at least 2 samples and at most 5 samples, in order to reduce false positives. To target for high-risk germline mutations, variants were required to co-segregate in at least one family.

Supplementary 3 Mutation enrichment analysis on protein-protein interaction interface during different stages of GeMSTONE pipeline

3a. GeMSTONE's performance on the ALS study showed 122 candidate variants with a significantly increased enrichment on protein-protein interaction (PPI) interface domains and residues derived from co-crystal structures[3, 4]. This trend is expected as a positive control based on disease mutation enrichment on PPI interfaces[3, 4]. 54,668 HGMD disease mutations (green) were plotted alongside GeMSTONE filtering results (blue), with case-only variants and naïve filtering presented as a negative control (grey).

3b. The same trend holds for GeMSTONE performance on the CRC study, where F1, F2, F3 and F5 represent variants kept after different sets of filters were applied. Each F-label corresponds to the first, second, third and fifth blocks of filters from **Supplementary Figure 1** respectively.

Supplementary 4 GeMSTONE visualization page

Through the results page of the GeMSTONE portal, users can visualize their variant statistics. The interactive menu atop the page shows the number of variants by chromosomal region, with variant quality, allele frequency (in ExAC), read depth, and insertion deletion length histograms, as well as tstv ratio and variant type comparisons. By clicking on a specific chromosome or using the filter toggle on the top left, the user can interactively explore these statistics at different resolution levels.

Supplementary Note 1

Overall, most tools we compared GeMSTONE to accept VCF and pedigree files as inputs, and can perform routine filtering on quality control and variant consequence (**Figure 1, Raw Data Input and Prioritization**). However, GeMSTONE stands out as a more comprehensive tool by including annotations at the variant, gene, pathway, and network level (**Figure 1, Knowledge-based Annotation and Prioritization**) and flexible co-segregation analysis using different inheritance models for potential germline mutation prioritization (**Figure 1, Inheritance Models**).

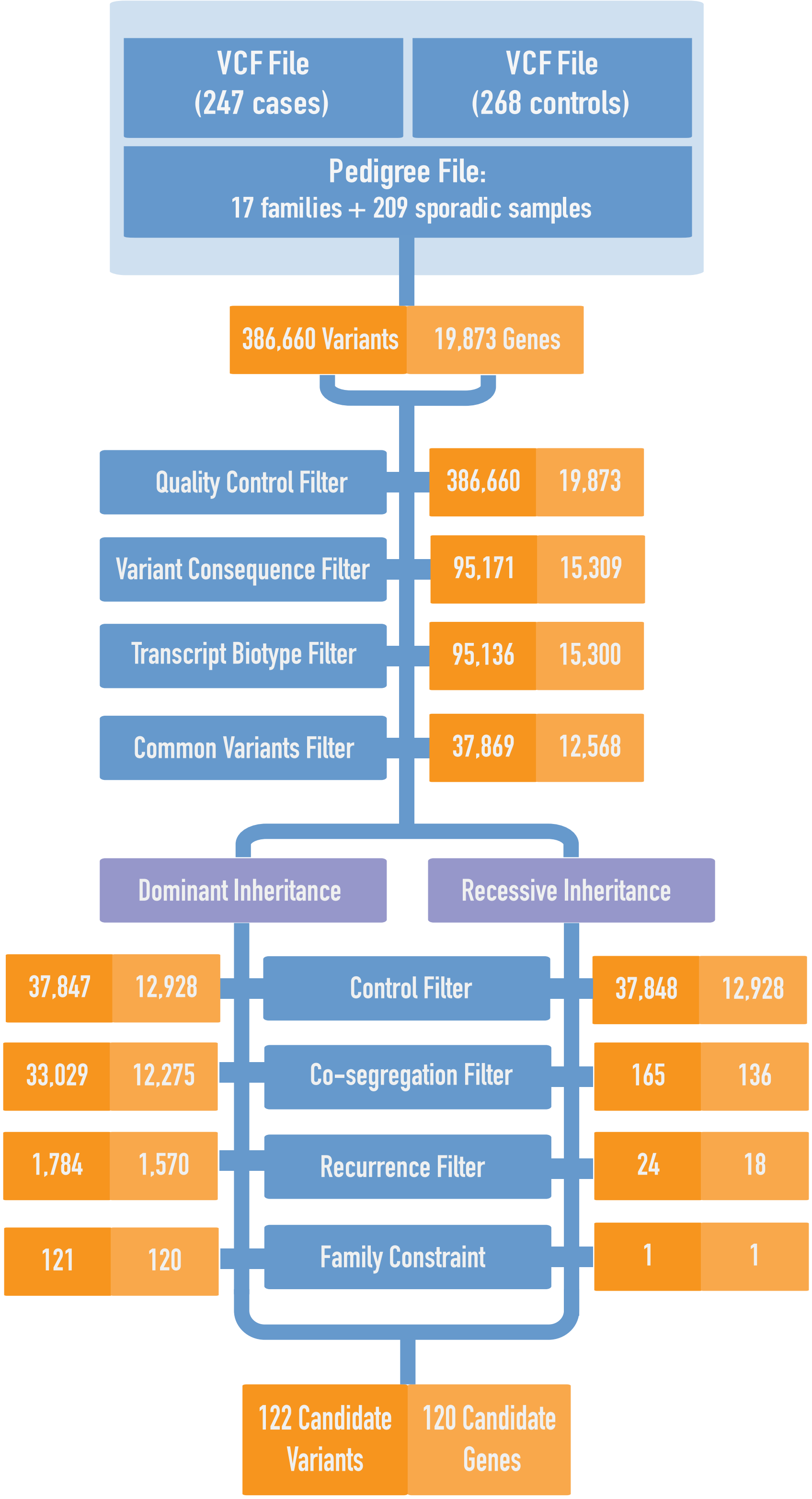
A keystone of GeMSTONE is the 'recipe' file (**Figure 1, Reproducibility**), which records all workflow parameters in a single file that can be shared and uploaded onto the site to reproduce a previous run. The recipe file can be used to (1) replicate results by rerunning the same workflow on the same dataset, (2) process new data with a known workflow or (3) modify parameters in a known workflow to evaluate study design. This approach has the potential to bring more transparency and openness to the bioinformatics community by enhancing the reproducibility of large-scale genomic studies.

References

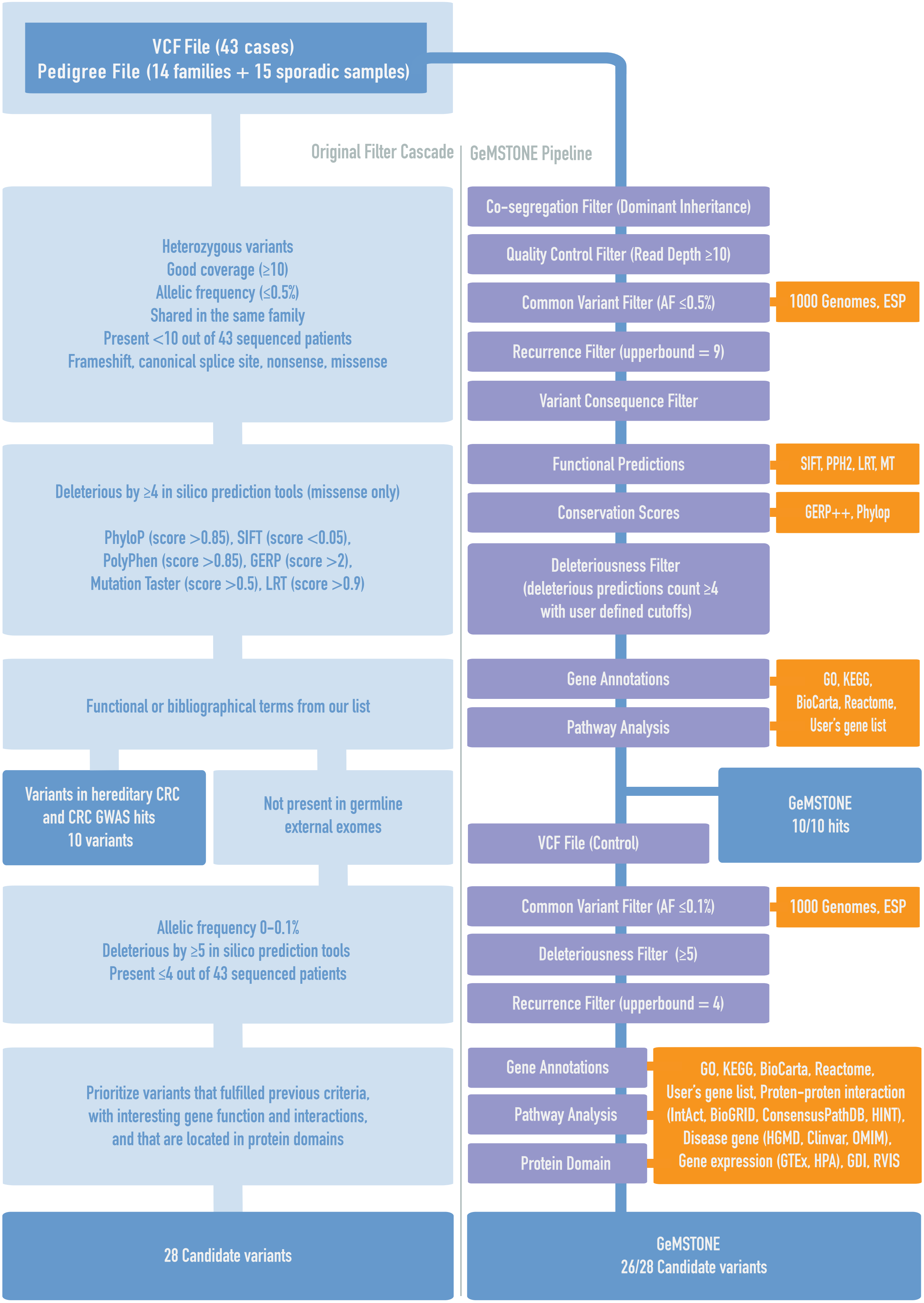
1. Esteban-Jurado, C. et al. *Genet Med* **17**, 131-142 (2015).
2. Cirulli, E.T. et al. *Science* **347**, 1436-1441 (2015).
3. Wang, X. et al. *Nat Biotechnol* **30**, 159-164 (2012).
4. Das, J. et al. *Hum Mutat* **35**, 585-593 (2014).
5. Aleman, A., Garcia-Garcia, F., Salavert, F., Medina, I. & Dopazo, J. *Nucleic Acids Res* **42**, W88-93 (2014).
6. Robinson, P.N. et al. *Genome Res* **24**, 340-348 (2014).
7. Smedley, D. et al. *Bioinformatics* **30**, 3215-3222 (2014).
8. Li, M.J. et al. *Hum Mutat* **36**, 496-503 (2015).
9. Ge, D. et al. *Bioinformatics* **27**, 1998-2000 (2011).
10. Sifrim, A. et al. *Genome Med* **4**, 73 (2012).
11. Cheng, Y.C. et al. *Nucleic Acids Res* **40**, W76-81 (2012).
12. Shetty, A.C. et al. *BMC Bioinformatics* **11**, 471 (2010).

Supplementary Figures

Supplementary 1 Recapitulation of a published CRC study

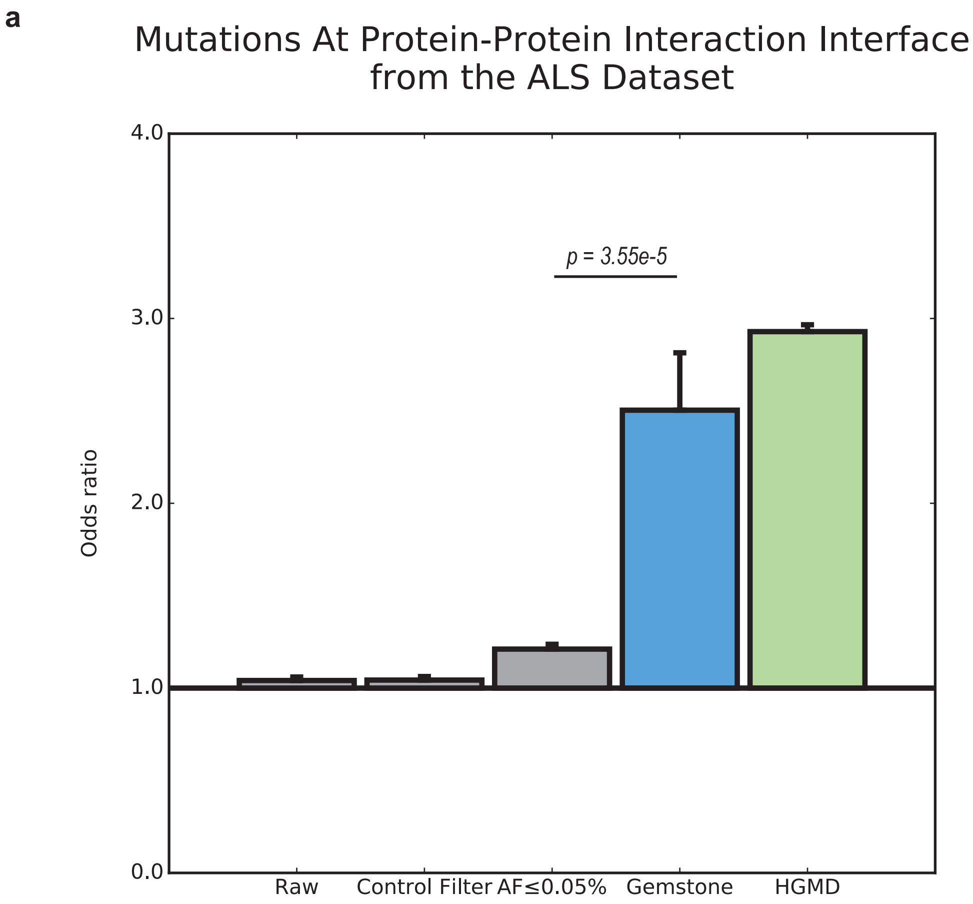


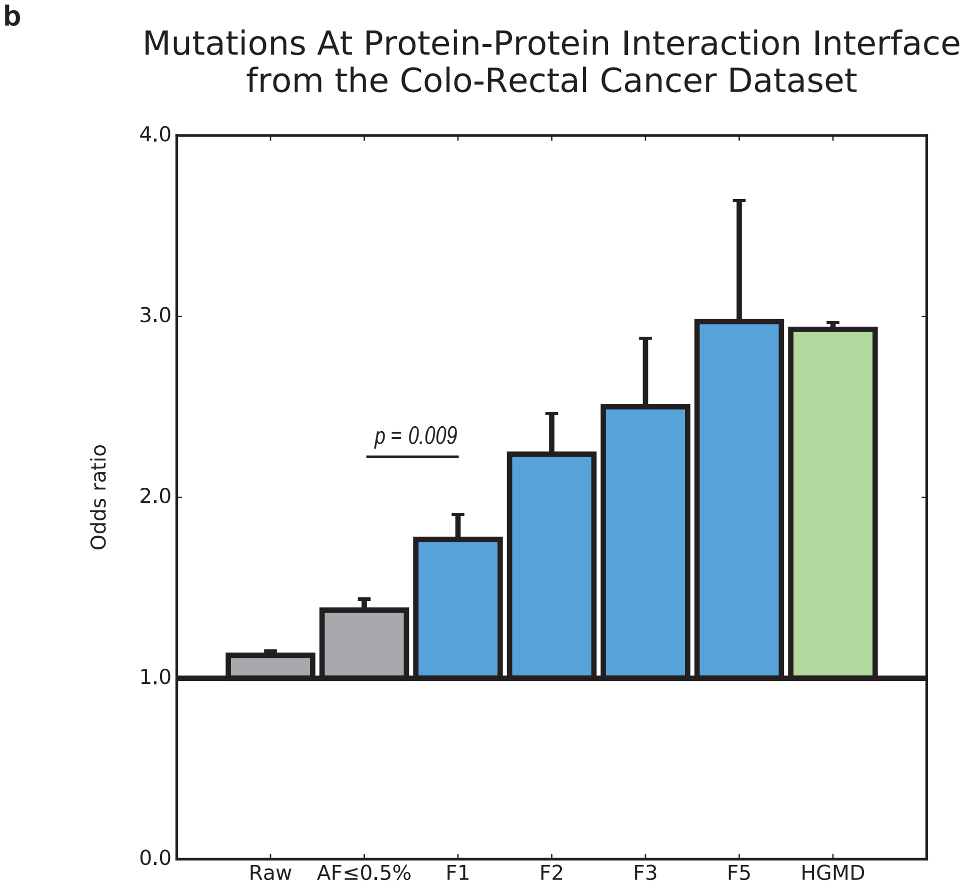
Supplementary 2 ALS study workflow



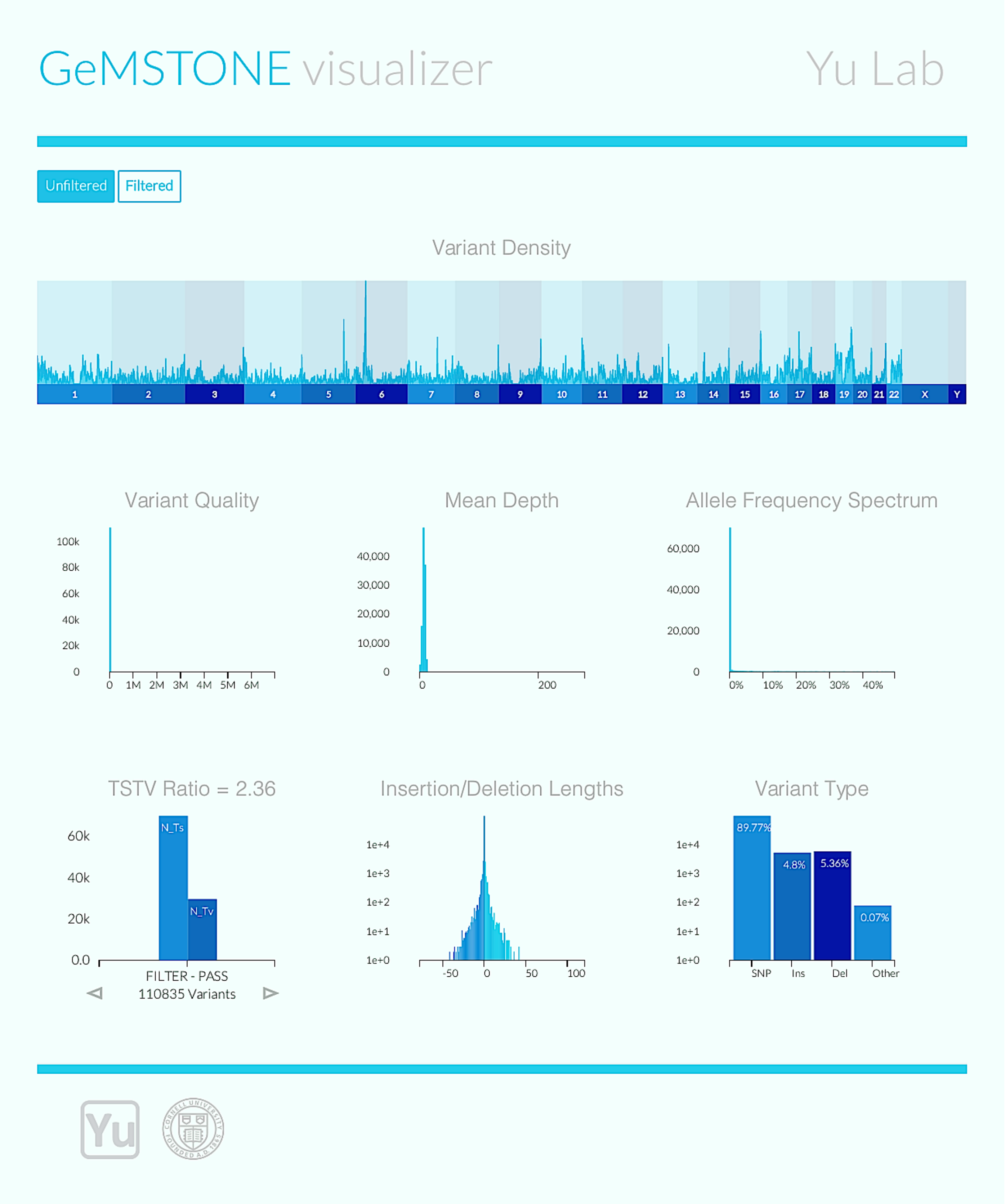
VCF File (43 cases)
Pedigree File (14 families + 15 sporadic samples)
Original Filter Cascade
GeMSTONE Pipeline
Heterozygous variants
Good coverage (≥10)
Allelic frequency (≤0.5%)
Shared in the same family
Present <10 out of 43 sequenced patients
Frameshift, canonical splice site, nonsense, missense
Deleterious by ≥4 in silico prediction tools (missense only)
PhyloP (score >0.85), SIFT (score <0.05),
PolyPhen (score >0.85), GERP (score >2),
Mutation Taster (score >0.5), LRT (score >0.9)
Functional or bibliographical terms from our list
Variants in hereditary CRC and CRC GWAS hits 10 variants
Not present in germline external exomes
Allelic frequency 0–0.1%
Deleterious by ≥5 in silico prediction tools
Present ≤4 out of 43 sequenced patients
Prioritize variants that fulfilled previous criteria,
with interesting gene function and interactions,
and that are located in protein domains
28 Candidate variants
Co-segregation Filter (Dominant Inheritance)
Quality Control Filter (Read Depth ≥10)
Common Variant Filter (AF ≤0.5%)
1000 Genomes, ESP
Recurrence Filter (upperbound = 9)
Variant Consequence Filter
Functional Predictions
SIFT, PPH2, LRT, MT
Conservation Scores
GERP++, Phylop
Deleteriousness Filter
(deleterious predictions count ≥4
with user defined cutoffs)
Gene Annotations
Pathway Analysis
GO, KEGG, BioCarta, Reactome, User's gene list
GeMSTONE 10/10 hits
VCF File (Control)
Common Variant Filter (AF ≤0.1%)
1000 Genomes, ESP
Deleteriousness Filter (≥5)
Recurrence Filter (upperbound = 4)
Gene Annotations
Pathway Analysis
Protein Domain
GO, KEGG, BioCarta, Reactome,
User's gene list, Proten-proten interaction
(IntAct, BioGRID, ConsensusPathDB, HINT),
Disease gene (HGMD, Clinvar, OMIM),
Gene expression (GTEx, HPA), GDI, RVIS
GeMSTONE 26/28 Candidate variants

Supplementary 3 Mutation enrichment on protein interface during different stages of GeMSTONE pipeline





Supplementary 4 GeMSTONE visualization page


GeMSTONE visualizer
Yu Lab
Unfiltered
Filtered
Variant Density
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y
Variant Quality
100k
80k
60k
40k
20k
0
0
1M
2M
3M
4M
5M
6M
Mean Depth
40,000
30,000
20,000
10,000
0
0
200
Allele Frequency Spectrum
60,000
40,000
20,000
0
0%
10%
20%
30%
40%
TSTV Ratio = 2.36
60k
40k
20k
0.0
N_Ts
N_Tv
FILTER - PASS
110835 Variants
◁
▷
Insertion/Deletion Lengths
1e+4
1e+3
1e+2
1e+1
1e+0
-50
0
50
100
Variant Type
1e+4
1e+3
1e+2
1e+1
1e+0
89.77%
4.8%
5.36%
0.07%
SNP
Ins
Del
Other
Yu