

Supplementary Material

Leveraging sequences missing from the human genome to diagnose cancer

Ilias Georgakopoulos-Soares^{1,2*}, Ofer Yizhar Barnea^{1,2*}, Ioannis Mouratidis³, Rachael Bradley^{1,2}, Ryder Easterlin^{1,2}, Candace Chan^{1,2}, Emmalyn Chen⁴, John S. Witte^{4,5,6}, Martin Hemberg^{7,8^}, Nadav Ahituv^{1,2^}

¹Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA.

²Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA.

³Department of Computer Science, Katholieke Universiteit Leuven, Leuven, Belgium.

⁴Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, USA

⁵Department of Urology, University of California San Francisco, San Francisco, California, USA

⁶Departments of Epidemiology and Population Health and Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA

⁷Wellcome Sanger Institute, Hinxton, UK.

⁸Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, USA

*These authors contributed equally to the work.

^Corresponding authors: mhemberg@bwh.harvard.edu, nadav.ahituv@ucsf.edu

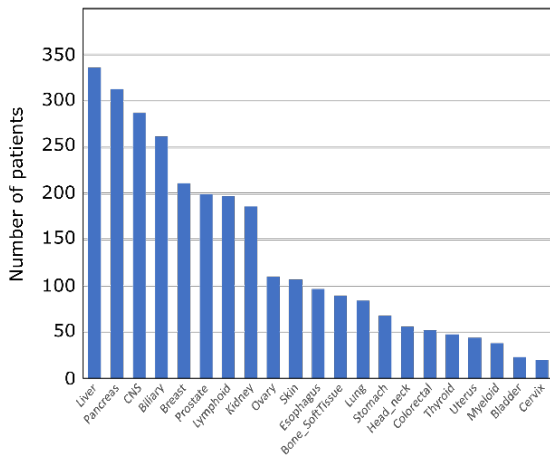
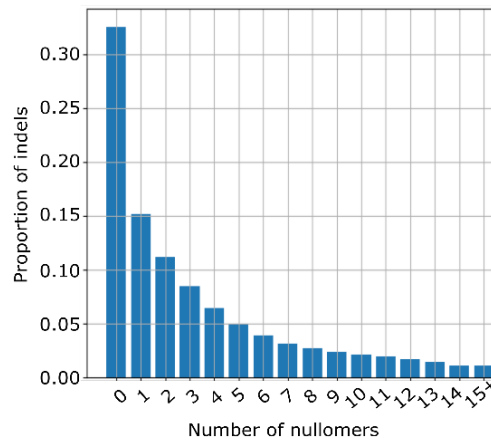
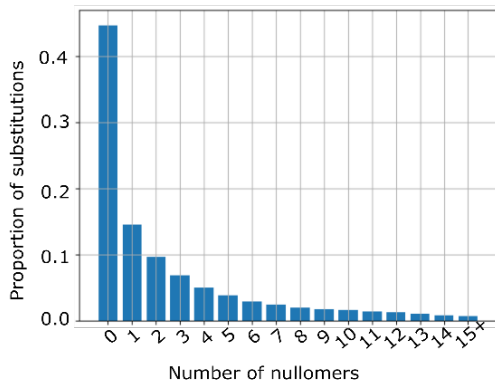
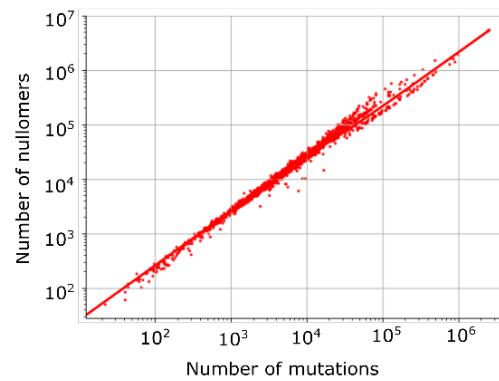
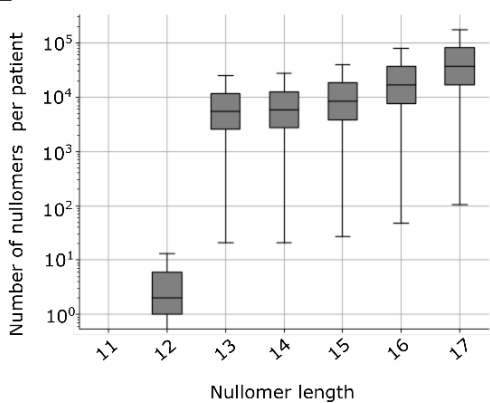
A**B****C****D****E**

Figure S1. (A) Number of patients per cancer tissue. (B-C) Number of nullomers resurfaced due to indels or substitutions observed for each tumor sample (per patient). (D) Association between number of mutations and number of nullomers observed. (E) Number of nullomers of different lengths observed per patient.

Table S1. Minimal recurrency thresholds and associated number of neomers per tissue type.

| Cancer tissue | Recurrency threshold | Number of neomers | Number of patients |
|----------------------|-----------------------------|--------------------------|---------------------------|
| Biliary | 2 | 166 | 262 |
| Bladder | 2 | 3,556 | 23 |
| Bone_SoftTissue | 2 | 2,018 | 89 |
| Breast | 2 | 27,280 | 211 |
| CNS | 2 | 7,238 | 287 |
| Cervix | 2 | 172 | 20 |
| Colorectal | 4 | 4,526 | 52 |
| Esophagus | 5 | 16,278 | 97 |
| Head_neck | 2 | 10,194 | 56 |
| Kidney | 2 | 17,584 | 186 |
| Liver | 3 | 4,756 | 336 |
| Lung | 3 | 3,028 | 84 |
| Lymphoid | 4 | 12,441 | 197 |
| Myeloid | 2 | 28 | 38 |
| Ovary | 2 | 13,744 | 110 |
| Pancreas | 2 | 61,557 | 313 |
| Prostate | 2 | 5,270 | 199 |
| Skin | 7 | 1,534 | 107 |
| Stomach | 3 | 1,832 | 68 |
| Thyroid | 2 | 60 | 48 |
| Uterus | 2 | 8,008 | 44 |

Table S2. Neomers resulting in neoantigens.

| Gene | Mutation | Number of Patients | Validation | Neomers |
|--------|----------|--------------------|-----------------------|--|
| EP300 | D1399N | 3 | Positive | AACACTATTGAGGTAA, ACTATTGAGGTAAGAT,ATCTTACCTCAATAGT, GAACACTATTGAGGTA,TACCTCAATAGTGTTT, TTACCTCAATAGTGTT |
| BRAF | L597Q | 2 | Positive | AGGTGATTTTGGTCAA,ATTTTGGTCAAGCTAC,GTAGCTTGACAAAAAT,TAGGTGATTTTGGTCA,TGACCAAAATCACC TA,TGTAGCTTGACCAAAA,TTGACCAAAATCACCT,TTTTGGTCAAGCTACA |
| TP53 | H193R | 8 | Positive | AAGACGCTGAGGAGGG,ACGCTGAGGAGGGGCC,ACTCGGATAAAGACGCT,AGACGCTGAGGAGGGG,AGCGTCTTAT CCGAGT,CACTCGGATAAAGACGC,CAGCGTCTTATCCGAG,CCACTCGGATAAAGACG,CCCCTCCTCAGCGTCT,CCCTC CTCAGCGTCTT,CCTCAGCGTCTTATCC,CGGATAAAGACGCTGAG,CGTCTTATCCGAGTGG,CTCAGCGTCTTATCCG,C TCGGATAAAGACGCTG,GCGTCTTATCCGAGTG,GGATAAAGACGCTGAGG,GGCCCTCCTCAGCGT,GTCTTATCCGAG TGGA,TCAGCGTCTTATCCGA,TCCACTCGGATAAAGAC, TCGGATAAAGACGCTGA |
| CTNNB1 | G34R | 3 | Positive | AATGGATTTCGAGAGTC,ACCAGAATGGATTTCGA,ACCTGGACTCTCGAAT,ATTCGAGAGTCCAGGT,CACCAGAATGGA TTCG,CCAGAATGGATTTCGAG,CCTGGACTCTCGAATC,CGAATCCATTCTGGTG,CGAGAGTCCAGGTAAG,CTCGAATC CATTCTGG,CTTACCTGGACTCTCG,GACTCTCGAATCCATT,GATTTCGAGAGTCCAGG,TACCTGGACTCTCGAA,TCGA ATCCATTCTGGT,TCGAGAGTCCAGGTA,TTACCTGGACTCTCGA,TTTCGAGAGTCCAGGTA |
| TP53 | G266V | 4 | Positive-High | AAGCTGTTCCGTACC,AAGCTGTTCCGTACCA,AATCTACTGGTACGGA,ACCAGTAGATTACCAC,ACTGGTACGGAACA GC,AGCTGTTCCGTACCAG,ATCTACTGGTACGGAA,CAAAGCTGTTCCGTAC,CCGTACCAGTAGATTA,CGTACCAGTA GATTAC,CTACTGGTACGGAACA,CTGGTACGGAACAGCT,CTGTTCCGTACCAGTA,GCTGTTCCGTACCAGT,GGTAAT CTACTGGTAC,GGTACGGAACAGCTTT,GTAATCTACTGGTACG,GTACCAGTAGATTACC,GTACGGAACAGCTTTG,GT GGTAATCTACTGGT,GTTCCGTACCAGTAGA,TAATCTACTGGTACGG,TACGGAACAGCTTTGA,TACTGGTACGGAACA G,TCAAAGCTGTTCCGTA,TCCGTACCAGTAGATT,TCTACTGGTACGGAAC,TGGTACGGAACAGCTT,TGTTCCGTACC AGTAG,TTCCGTACCAGTAGAT |
| TP53 | R337C | 7 | Positive-High | AACATCTCGAAGCACT,AAGCACTCACGCCAC,ACATCTCGAAGCACTC,ACTCACGCCACGGAT,AGCACTCACGCC CAGC,AGTGCTTCGAGATGTT,ATCCGTGGGCGTGAGT,ATCTCGAAGCACTCAC,CACTCACGCCACGGA,CATCTCG AAGCACTCA,CCGTGGGCGTGAGTGC,CGTGAGTGCTTCGAGA,CGTGGGCGTGAGTGCT,CTCGAAGCACTCACGC,G AACATCTCGAAGCAC,GAAGCACTCACGCCCA,GAGTGCTTCGAGATGT,GCACTCACGCCACGG,CGGTGAGTGCTT CGAG,GGAACATCTCGAAGCA,GGCGTGAGTGCTTCGA,GTGAGTGCTTCGAGAT,GTGCTTCGAGATGTT,GTGGGCG TGAGTGCTT,TCCGTGGGCGTGAGTG,TCGAAGCACTCACGCC,TCTCGAAGCACTCACG,TCGAGTGCTTCGAGATG,TG CTTCGAGATGTTCC,TGGGCGTGAGTGCTT |
| TP53 | C141Y | 4 | Positive-Intermediate | AAGACCTACCTGTGC,CAGGGTAGGTCTTGGC,CTGCACAGGGTAGGTC,GACCTACCTGTGCAG,GCACAGGGTAG GTCTT,GCCAAGACCTACCTG |
| TP53 | C135F | 4 | Positive-High | GTCTTGGCCAGTTGGA,TCCAAGTGGCCAAGAC |
| ERBB2 | D769Y | 3 | Positive-Intermediate | AAATCTTATACGTAAG,AAGAAATCTTATACGT,AATCTTATACGTAAGC,ACGTATAAGATTTCTT,AGAAATCTTATACGT A,AGGGGCTTACGTATAA,ATACGTAAGCCCTCC,ATCTTATACGTAAGCC,CTTACGTATAAGATTT,CTTATACGTAAG CCCC,GAAATCTTATACGTAA,GAGGGGCTTACGTATA,GCTTACGTATAAGATT,GGAGGGGCTTACGTAT, GGCTTACGTATAAGAT,GGGCTTACGTATAAGA,GGGGCTTACGTATAAG, TACGTAAGCCCTCCA,TACGTATAAGATTTCT,TATACGTAAGCCCTC,TCTTATACGTAAGCCC,TGGAGGGGCTTAC GTA,TTACGTATAAGATTTCT,TTATACGTAAGCCCT |
| TPR | S2155L | 2 | Positive | AACCTGCGGCAAACCTA,ACCTGCGGCAAACCTAA,AGCAACCTGCGGCAAA,ATTTAGTTTGGCCGAG,CATTTAGTTTGGC GCA,CCATTTAGTTTGGCCG,CCTGCGGCAAACCTAAA,CGGCAAACCTAAATGGA,CTGCGGCAAACCTAAAT,GCAACCTGC GGCAAAC,GCGGCAAACCTAAATGG,GGCAAACCTAAATGGAC,GTCCATTTAGTTTGGC,GTTTGCGGCAAGTTGC,TAGTT TGCCGCAAGTT,TCCATTTAGTTTGGCC,TGCGGCAAACCTAAATG,TTAGTTTGGCCGAGGT,TTAGTTTGGCCGAGG,TT TGCCGCAAGTTGCT |
| ASH2L | E617K | 2 | Positive | ATCACGTGGAGACAAA,TATCACGTGGAGACAA,TTGTCTCCACGTGATA,TTGTCTCCACGTGAT |
| ANKLE2 | E103Q | 2 | Positive | CAGAAAAAATTTGGCTC,GAGCCAATTTTTCTG |
| DNMT1 | E432K | 2 | Positive | ATCGAGAAGAATATCA,CGAGAAGAATATCAAA,GAAGAGTTTGATATTC,GAATATCAAACCTTTC,TCGAGAAGAATATC AA,TGATATTTCTCTCGAT,TTGATATTTCTCTCGA,TTTGATATTTCTCTCG |

Table S3. Promoter sequences cloned into pGL promoter luciferase plasmid.

| | | |
|---------------------------------------|---|-------------|
| <p><i>FRS3-PRICKL4</i> (REF)</p> | <p>CGCTCCCGGGTCCCCTGCAGCAGCCGCGCCCGCCCGGAGCTAAGGCCTCCCGCCCCGCCCCG CCCCGCGGCCGCGGGTTCGAGGTCACCGGCAGCACCAACGAGACGCTGCGGGCTGGCGGACCCG AGCGGGGGCGACCCGGCACGCAGCCTAGAAGTCCCTTTGGAGGACTTTGCAGTCGGGAGCAGTGAG CATGCGCAGACGCCAGAGGGCGCTCCTGGCCAGGGGCTGAGCAAGCCCGGAGGGGACAGGGCGGA GCCTGCGGCGGAGGCGGGGCGCGGAGGCGGGGCTGAGGACACCAAAGTTGGGGGTGGTCCGGGG AGCGGAAAGGGGACGGAGGGTTGGGCGAGAAACGAAGGGAAGGGGAGAGGGTCTGAAGGGACCTG AAAGGAGAAGGGCAAGACGATGGAACCTGGAAGAGGGTTGAGGGGAGGGGAAGGGGCCAGGGGGA CCACTGGGGACAAGGGGGAGAGAAGTGAAGGGAACACGCGGAGGAGCGGGGCTGGGAGACCACA GGGAGAGTTTGGGGAGACCTGGGAAGTGCAGGAGACAGGGCTGCGGGCCCCGGCGAGGCGTTTCC ATTCAGCGAATGTTGAAAACCCAGCGTAAACGCGGCTCCCGCGTCTGCTGGATGCCTGCCGGGCG CAGTCCGAGCCAGAGAGCCCTTGCACCTGCATGGCAGAGGCCATGGAGCCCGCACATTTCCACCCTG AGATCCATAAATGTGGTTTTGTTTTCTTTTTTTTTCTCCCCCAATCTGCAGCTCCAGGCCCTCC CTGAGCCGCCCTCTCCACCAGCCGGGGAGGGGCGGGCGGGGGCCGCTGGGGCGTGGGAGAGGA AGGGCCGCGCTATTAGGGCGGGCGGGGCGGACGGGGCGGGCGCTGAAGCCTTGGGGTGGGGA CCGGTTAGGGGCTGACGGCAGAGTCCGTCCAGAACTGCCCGGACAGCGACGACGAGGGT GAGACCCAGCGGAGTCTGGTCCCCTCCCCTAAAATCCCAGCAGCCAGACCGGG</p> | <p>1035</p> |
| <p><i>FRS3-PRICKL4</i> (MUT)</p> | <p>CGCTCCCGGGTCCCCTGCAGCAGCCGCGCCCGCCCGGAGCTAAGGCCTCCCGCCCCGCCCCG CCCCGCGGCCGCGGGTTCGAGGTCACCGGCAGCACCAACGAGACGCTGCGGGCTGGCGGACCCG AGCGGGGGCGACCCGGCACGCAGCCTAGAAGTCCCTTTGGAGGACTTTGCAGTCGGGAGCAGTGAG CATGCGCAGACGCCAGAGGGCGCTCCTGGCCAGGGGCTGAGCAAGCCCGGAGGGGACAGGGCGGA GCCTGCGGCGGAGGCGGGGCGCGGAGGCGGGGCTGAGGACACCAAAGTTGGGGGTGGTCCGGGG AGCGGAAAGGGGACGGAGGGTTGGGCGAGAAACGAAGGGAAGGGGAGAGGGTCTGAAGGGACCTG AAAGGAGAAGGGCAAGACGATGGAACCTGGAAGAGGGTTGAGGGGAGGGGAAGGGGCCAGGGGGA CCACTGGGGACAAGGGGGAGAGAAGTGAAGGGAACACGCGGAGGAGCGGGGCTGGGAGACCACA GGGAGAGTTTGGGGAGACCTGGGAAGTGCAGGAGACAGGGCTGCGGGCCCCGGCGAGGCGTTTCC ATTCAGCGAATGTTGAAAACCCAGCGTAAACGCGGCTCCCGCGTCTGCTGGATGCCTGCCGGGCG CAGTCCGAGCCAGAGAGCCCTTGCACCTGCATGGTAGAGGCCATGGAGCCCGCACATTTCCACCCTG AGATCCATAAATGTGGTTTTGTTTTCTTTTTTTTTCTCCCCCAATCTGCAGCTCCAGGCCCTCC CTGAGCCGCCCTCTCCACCAGCCGGGGAGGGGCGGGCGGGGGCCGCTGGGGCGTGGGAGAGGA AGGGCCGCGCTATTAGGGCGGGCGGGGCGGACGGGGCGGGCGCTGAAGCCTTGGGGTGGGGA CCGGTTAGGGGCTGACGGCAGAGTCCGTCCAGAACTGCCCGGACAGCGACGACGAGGGT GAGACCCAGCGGAGTCTGGTCCCCTCCCCTAAAATCCCAGCAGCCAGACCGGG</p> | <p>1035</p> |
| <p><i>TMEM127-CIAO1</i> (REF)</p> | <p>GGGGCCGAGCGTCCAGGCCCAATCCCGCAACGCCCGGACAGACCCGGGGCCGATGCACTTCCG GCTTCTCCTCCTCTGCCTCCGCCTCCTCCGGCTCCCGCCCCGTCCCGCGCGCGTACAGCCTG GGCGCGCCCGCGGCTCTGCGCAGGGCCTCACGGGACTGGTAGTTGAGGCGGTCCCGCGATGCTTT CCCTGCCGAGGTGAGGGGAGCGACTACAACCTCCAGAGGGTATGGCGGCCGAGCAGTGGGCAGTT GAGTTTCCGGGGCGACCCAGTCTCCGCGGAAGCGGGAGCCTCTGTGGCCGCGGAAGCCTGGA GTGGGCGGTACGCAGACGC</p> | <p>350</p> |
| <p><i>TMEM127-CIAO1</i> (MUT)</p> | <p>GGGGCCGAGCGTCCAGGCCCAATCCCGCAACGCCCGGACAGACCCGGGGCCGATGCACTTCCG GCTTCTCCTCCTCTGCCTCCGCCTCCTCCGGCTCCCGCCCCGTCCCGCGCGCGTACAGCCTG GGCGCGCCCGCGGCTCTGCGCAGGGCCTCACGGGACTGGTAGTTGAGGCGGTCCCGCGATGCTTT CCCTGCCGAGGTGAGGGGAGCGACTACAACCTCCAGAGGGTATGGCGGCCGAGCAGTGGGCAGTT GAGTTTCCGGGGCGACCCAGTCTCCGCGGAAGCGGGAGCCTCTGTGGCCGCGGAAGCCTGGA GTGGGCGGTACGCAGACGC</p> | <p>350</p> |
| <p><i>LLRC28-TTC23</i> (REF)</p> | <p>CCGACTAAATTACCCTTCAAGACGCTAAGCAAAGCCACTCCGTCGACATCAAACCCACGCCGGAG CCGGAAGCAAGAGTAGAGCCGCGGGGCGGGCCAGCATAGAGGAAGTAGCGCTCTCGGCCAATACAC AGATCGCTTGCAAACGAGCATTTTCCCGCGCTATCGTGAAGCGCGAGAACACAGGGGATTATCG CGAGGTTCCGTGCTCCTCTCCCGCTCCCTCTTCCCAGCTCCCTCTGGCGCGGGTTCCGCTCCACCC CCGGAGCCCCGCCCCGACGCCGCGCTCCGCCCGCCCTCGCTGGCTCCCTGCCCGCCCCG CCCGCGTCCCCGCTT</p> | <p>350</p> |
| <p><i>LLRC28-TTC23</i> (MUT)</p> | <p>CCGACTAAATTACCCTTCAAGACGCTAAGCAAAGCCACTCCGTCGACATCAAACCCACGCCGGAG CCGGAAGCAAGAGTAGAGCCGCGGGGCGGGCCAGCATAGAGGAAGTAGCGCTCTCGGCCAATACAC AGATCGCTTGCAAACGAGCATTTTCCCGCGCTATCGTGAAGCGCGAGAACACAGGGGATTATCG CGAGGTTCCGTGCTCCTCTCCCGCTCCCTCTTCCCAGCTCCCTCTGGCGCGGGTTCCGCTCCACCC CCGGAGCCCCGCCCCGACGCCGCGCTCCGCCCGCCCTCGCTGGCTCCCTGCCCGCCCCG CCCGCGTCCCCGCTT</p> | <p>350</p> |
| <p><i>RPS2-SNHG9</i> (REF)</p> | <p>GGGAGCCTAGACCCGACCCGATGTCCGCGGATTCCCGCCGCCACGCAGAGGCCGCTGCAGCGAC CAACAGGACTCACGTGTTTTGTCGAAAAGAGAACGAGACCTACTGGGAAGCAGCTTTTATAGCAGC CCAAGCGCCGAGATCTCCGACGCCCGCCCCAAGCGGGAGCGGGCCGAGCTCCTATAAGACAAC CTGTGATTGGCTCCGCGTGCCCGCCCTCACCGGCTGTGAGTGCTTTGCCCGTCCGGCCCCAG</p> | <p></p> |

| | | |
|------------------------|---|-----|
| | CCGCGGCCCGGGAATCTACGTCACCCGAAAAGCGACTATAAACGCCGGCGCCTCCGTCCCCAGCCG CGGCTCGGGAATCCACCC | |
| <i>RPS2-SNHG9</i> | GGGAGCCTAGACCCGACCCGATGTCCGCGGATTCCCGCCGCCACGCAGAGGCCCGCTGCAGCAAC CAACAGGACTCACGTGTTTTGTCGGAAAAGAAGAACGAGACCTACTGGGAAGCAGCTTTTATAGCAG CCAAGCGCCGCGAGATCTCCGCAGCCCCGCCCAAGCGGGAGCGGGCCGAGCTCCTATAAGACAAC CTGTGATTGGCTCCGCGGTGCCCGCCCTACCGGGCTCTGAGTGCTCTTGCCCGTCCGGCCCCAG CCGCGGCCCGGGAATCTACGTCACCCGAAAAGCGACTATAAACGCCGGCGCCTCCGTCCCCAGCCG CGGCTCGGGAATCCACCC | |
| <i>GNAI2 (REF)</i> | CATGTATGTCCATTATGGGCCATCTGTTGACTGTCCCTGCGGACCCGGCCTCCGGGGCACAGCCTTG GAGGTTTCAAATGGCATCCGCTTCGCTGCTGCCTCAGGGAGGAAGAGAAGGGGCGGGAAGGGGCG GCGGCCAGCCTCCTGGGCTGCTTTGCGTTGGGCGCTTGCCCTTCCTGTCTCTGCTGTGAGTGCGG CAGCGGCAGTGAGTGGGTGTCGACGCGGCGGAATGCCCGTCGCTGCTGCTGCTGCTGCCCGACGGG CCTGGGGAGGGCACTCCGGTACCGCTGCCTCCAGTGCCCAGCCAGGAGTCTCTAGACAGCCTCTC CCCAGGCGATTGCAGTT | 350 |
| <i>GNAI2 (MUT)</i> | CATGTATGTCCATTATGGGCCATCTGTTGACTGTCCCTGCGGACCCGGCCTCCGGGGCACAGCCTTG GAGGTTTCAAATGGCATCCGCTTCGCTGCTGCCTCAGGGAGGAAGAGAAGGGGCGGGAAGGGGCG GCGGCCAGCCTCCTGGGCTGCTTTGCGTTGGGCGCTTGCCCTTCCTGTCTCTGCTGTGAGTGCGG CAGCGGCAGTGAGTGGGTGTCGACGCGGCGGAATGCCCGTCGCTGCTGCTGCTGCTGCCCGACGGG CCTGGGGAGGGCACTCCGGTACCGCTGCCTCCAGTGCCCAGCCAGGAGTCTCTAGACAGCCTCTC CCCAGGCGATTGCAGTT | 350 |