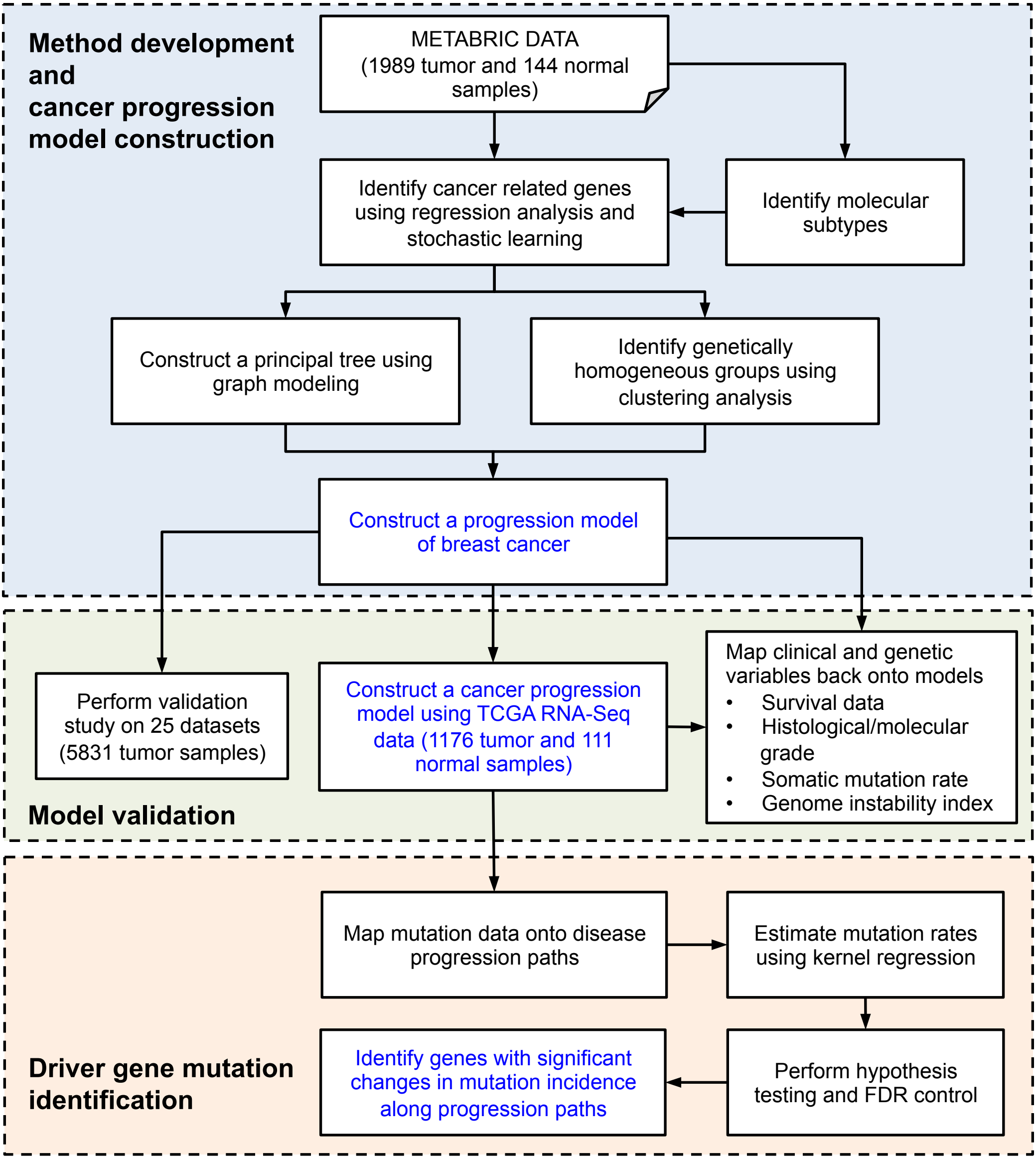
Method development and cancer progression model construction
METABRIC DATA (1989 tumor and 144 normal samples)
Identify cancer related genes using regression analysis and stochastic learning
Identify molecular subtypes
Construct a principal tree using graph modeling
Identify genetically homogeneous groups using clustering analysis
Construct a progression model of breast cancer
Perform validation study on 25 datasets (5831 tumor samples)
Construct a cancer progression model using TCGA RNA-Seq data (1176 tumor and 111 normal samples)
Map clinical and genetic variables back onto models
• Survival data
• Histological/molecular grade
• Somatic mutation rate
• Genome instability index
Model validation
Map mutation data onto disease progression paths
Estimate mutation rates using kernel regression
Perform hypothesis testing and FDR control
Identify genes with significant changes in mutation incidence along progression paths
Driver gene mutation identification

Figure 1: Overview of the presented stepwise study consisting of three major components. First, a comprehensive bioinformatics pipeline was developed and a progression model of breast cancer was constructed. Then, a large-scale validation study was performed to evaluate the validity of the constructed model. Finally, a cancer genome analysis, focusing primarily on the detection of cancer driver gene mutations, was conducted that demonstrated the utility of the progression model. The study incorporated extensive algorithm development (**Online Methods**) and the analysis of 27 breast cancer datasets for model construction and validation (**Online Methods – Datasets**)

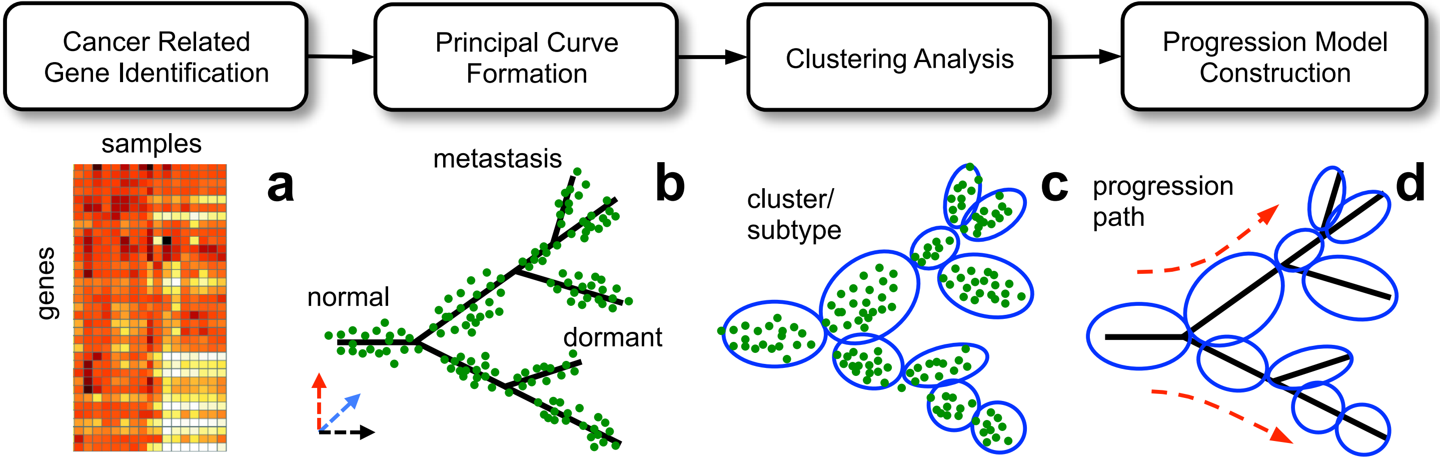

Figure 2: Overview of the bioinformatics pipeline for cancer progression modeling.

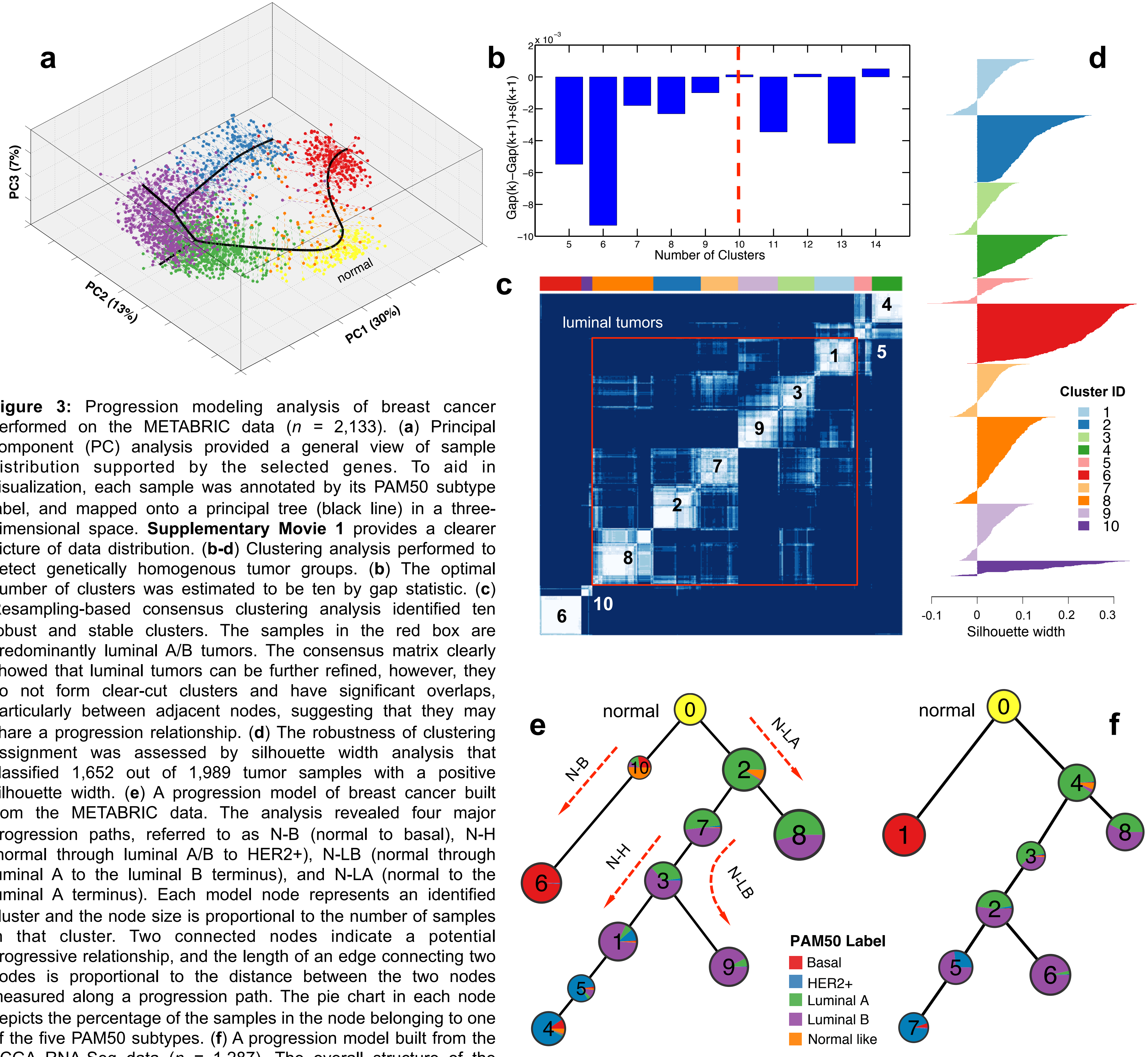

Figure 3: Progression modeling analysis of breast cancer performed on the METABRIC data (n = 2,133). (**a**) Principal component (PC) analysis provided a general view of sample distribution supported by the selected genes. To aid in visualization, each sample was annotated by its PAM50 subtype label, and mapped onto a principal tree (black line) in a three-dimensional space. **Supplementary Movie 1** provides a clearer picture of data distribution. (**b-d**) Clustering analysis performed to detect genetically homogenous tumor groups. (**b**) The optimal number of clusters was estimated to be ten by gap statistic. (**c**) Resampling-based consensus clustering analysis identified ten robust and stable clusters. The samples in the red box are predominantly luminal A/B tumors. The consensus matrix clearly showed that luminal tumors can be further refined, however, they do not form clear-cut clusters and have significant overlaps, particularly between adjacent nodes, suggesting that they may share a progression relationship. (**d**) The robustness of clustering assignment was assessed by silhouette width analysis that classified 1,652 out of 1,989 tumor samples with a positive silhouette width. (**e**) A progression model of breast cancer built from the METABRIC data. The analysis revealed four major progression paths, referred to as N-B (normal to basal), N-H (normal through luminal A/B to HER2+), N-LB (normal through luminal A to the luminal B terminus), and N-LA (normal to the luminal A terminus). Each model node represents an identified cluster and the node size is proportional to the number of samples in that cluster. Two connected nodes indicate a potential progressive relationship, and the length of an edge connecting two nodes is proportional to the distance between the two nodes measured along a progression path. The pie chart in each node depicts the percentage of the samples in the node belonging to one of the five PAM50 subtypes. (**f**) A progression model built from the TCGA RNA-Seq data (n = 1,287). The overall structure of the progression models constructed using the two independent datasets is almost identical.

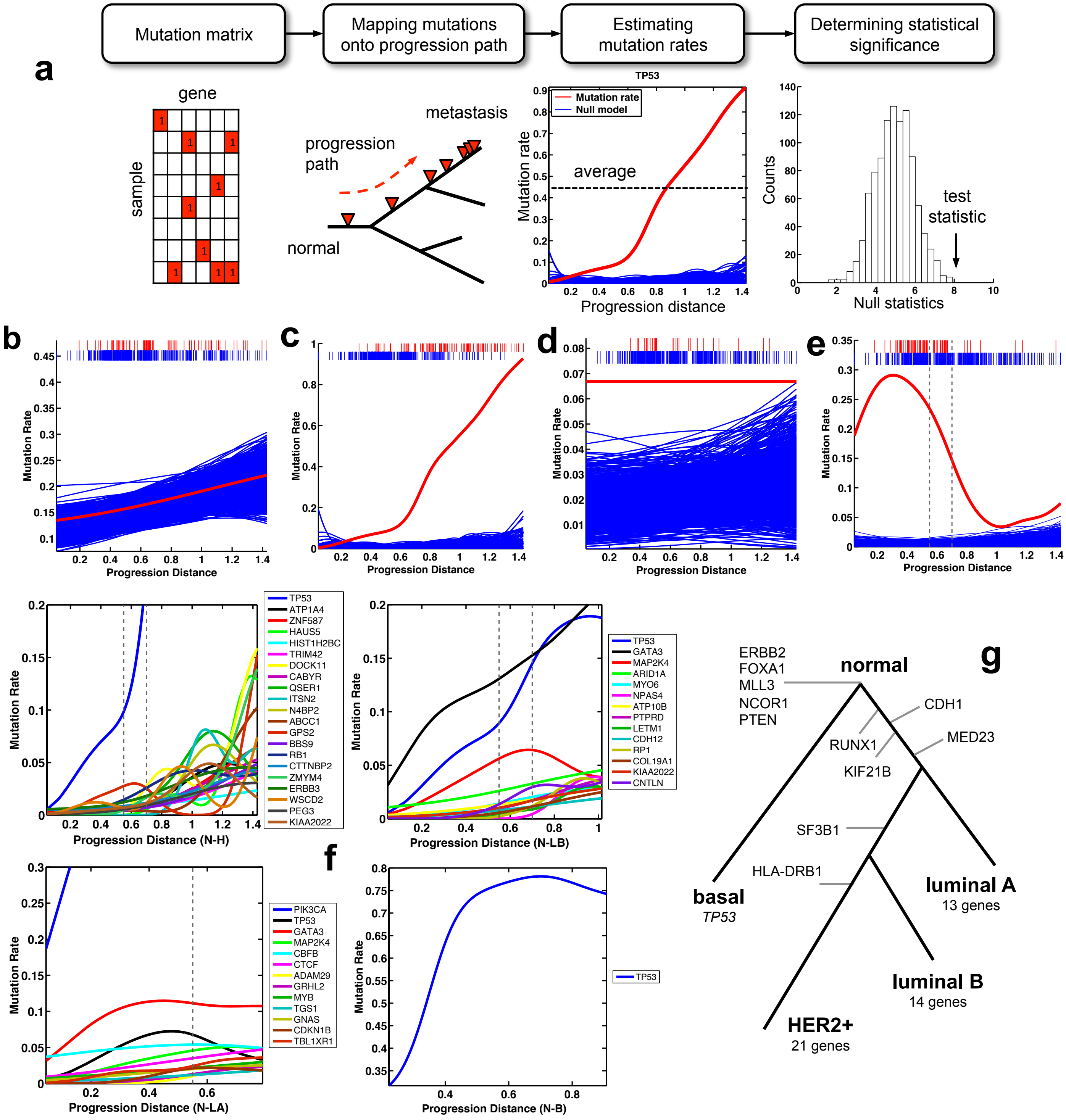

Figure 5: Pseudo-time series analysis performed on the TCGA mutation data (n = 958) to identify gene mutations associated with cancer progression. Fifty one genes were found to have significant changes in their mutation incidences along progression paths (FDR<0.05). (**a**) Overview of the proposed MutationPattern method used to delineate the dynamic patterns of individual gene mutations along a progression path. (**b-e**) Four distinct mutation patterns were observed. Examples of each are depicted: (**b**) *TTN*, (**c**) *TP53*, (**d**) *MLL3*, and (**e**) *CDH1*. The red line depicts the estimated mutation rate, and blue lines were generated from null models built by assuming that the corresponding gene plays no role in cancer development. Each red or blue line in the bar above the figure represents the presence or absence of a mutation in a sample, respectively. The first and second broken lines in (**e**) indicate the locations where the N-H path intersects with the LA terminal and LB terminal, respectively. (**f**) Genes showing an upward mutation trend along the N-LA, N-LB, N-H and N-B progression paths. (**g**) Mapping of significantly progression-associated genes onto the TCGA model. Genes reported at the end of a path are those with an upward trend along the entire path. Genes with a bell-shaped pattern are marked at the bell-peak location. Genes associated with normal samples are those mutated more frequently than random chance, but do not have significant changes along any progression path.

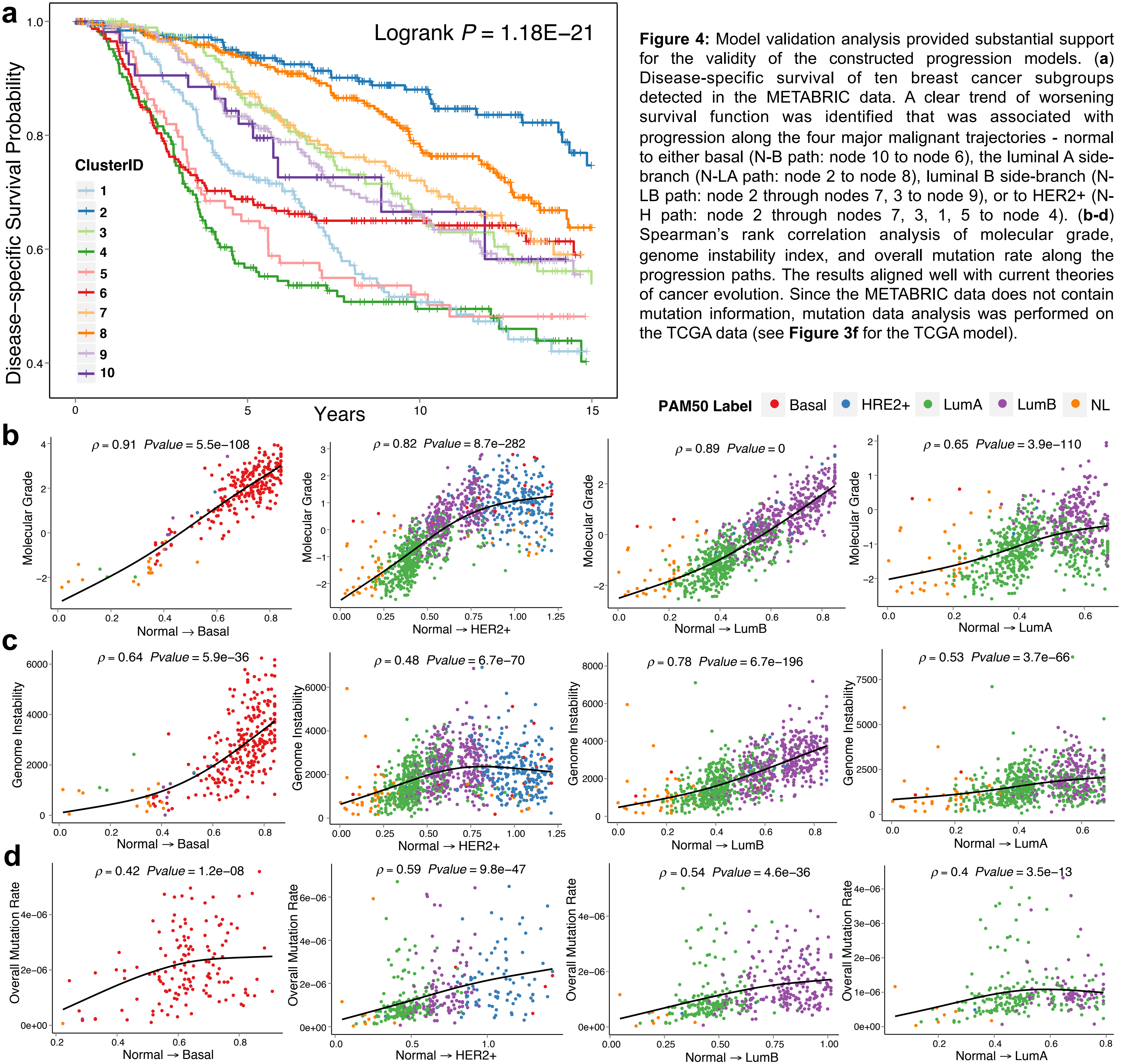

Figure 4: Model validation analysis provided substantial support for the validity of the constructed progression models. (**a**) Disease-specific survival of ten breast cancer subgroups detected in the METABRIC data. A clear trend of worsening survival function was identified that was associated with progression along the four major malignant trajectories - normal to either basal (N-B path: node 10 to node 6), the luminal A side-branch (N-LA path: node 2 to node 8), luminal B side-branch (N-LB path: node 2 through nodes 7, 3 to node 9), or to HER2+ (N-H path: node 2 through nodes 7, 3, 1, 5 to node 4). (**b-d**) Spearman's rank correlation analysis of molecular grade, genome instability index, and overall mutation rate along the progression paths. The results aligned well with current theories of cancer evolution. Since the METABRIC data does not contain mutation information, mutation data analysis was performed on the TCGA data (see **Figure 3f** for the TCGA model).

Table 1: Fifty one genes were identified to have a significant change in mutation incidence in at least one progression path (FDR < 0.05). If a gene was found in multiple paths, only the smallest P-value and FDR were reported. A path highlighted in red means that a gene has a monotonically increasing mutation pattern in that path, and a path highlighted in green means that a gene has a bell-shaped mutation pattern. # Samples: the number of samples with mutations.

Rank	Gene	Full Name	N-H	N-LB	N-LA	N-B	# Samples	P-value	FDR
1	PIK3CA	phosphoinositide-3-kinase, catalytic, alpha polypeptide	●🟩	●🟩	●🟥		314	0	0
2	TP53	tumor protein p53	●🟥	●🟥	●🟥	●🟥	291	0	0
3	CDH1	cadherin 1, type 1, E-cadherin (epithelial)	●🟩	●🟩	●🟩		103	0	0
4	GATA3	GATA binding protein 3	●🟩	●🟥	●🟥		95	0	0
5	MAP3K1	mitogen-activated protein kinase kinase kinase 1	●🟩	●🟩			70	0	0
6	MAP2K4	mitogen-activated protein kinase kinase 4	●🟩	●🟥	●🟥		32	0	0
7	RUNX1	runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene)	●🟩	●🟩	●🟩		28	0	0
8	TBX3	T-box 3 (ulnar mammary syndrome)	●🟩	●🟩	●🟩		27	0	0
9	CBFB	core-binding factor, beta subunit	●🟩	●🟩	●🟥		23	0	0
10	CTCF	CCCTC-binding factor (zinc finger protein)	●🟩	●🟩	●🟥		17	0	0
11	ATP1A4	ATPase, Na+/K+ transporting, alpha 4 polypeptide	●🟥				13	0	0
12	ZNF587	zinc finger protein 587	●🟥				10	0	0
13	ADAM29	ADAM metallopeptidase domain 29			●🟥		9	0	0
14	HLA-DRB1	major histocompatibility complex, class II, DR beta 1	●🟩				8	0	0
15	HAUS5	HAUS augmin-like complex, subunit 5	●🟥				5	0	0
16	ARID1A	AT rich interactive domain 1A (SWI-like)		●🟥			26	1.00E-04	2.80E-03
17	MYO6	myosin VI		●🟥			10	1.00E-04	2.80E-03
18	NPAS4	neuronal PAS domain protein 4		●🟥			9	1.00E-04	2.80E-03
19	MED23	mediator complex subunit 23		●🟩	●🟩		14	1.00E-04	3.03E-03
20	HIST1H2BC	histone cluster 1, H2bc	●🟥				6	1.00E-04	4.20E-03
21	TRIM42	tripartite motif-containing 42	●🟥				6	1.00E-04	4.20E-03
22	ATP10B	ATPase, class V, type 10B		●🟥			17	2.00E-04	5.17E-03
23	DOCK11	dedicator of cytokinesis 11	●🟥				20	2.00E-04	6.72E-03
24	CABYR	calcium binding tyrosine-(Y)-phosphorylation regulated	●🟥				6	2.00E-04	6.72E-03
25	GRHL2	grainyhead-like 2 (Drosophila)			●🟥		8	3.00E-04	7.26E-03
26	QSER1	glutamine and serine rich 1	●🟥				20	3.00E-04	7.96E-03
27	ITSN2	intersectin 2	●🟥				12	3.00E-04	7.96E-03
28	N4BP2	NEDD4 binding protein 2	●🟥				10	3.00E-04	7.96E-03
29	ABCC1	ATP-binding cassette, sub-family C (CFTR/MRP), member 1	●🟥				11	4.00E-04	9.60E-03
30	SF3B1	splicing factor 3b, subunit 1, 155kDa		●🟩	●🟩		16	5.00E-04	1.01E-02
31	MYB	v-myb myeloblastosis viral oncogene homolog (avian)			●🟥		12	5.00E-04	1.01E-02
32	GPS2	G protein pathway suppressor 2	●🟥				11	5.00E-04	1.10E-02
33	BBS9	Bardet-Biedl syndrome 9	●🟥				9	5.00E-04	1.10E-02
34	RB1	retinoblastoma 1 (including osteosarcoma)	●🟥				19	7.00E-04	1.41E-02
35	CTTNBP2	cortactin binding protein 2	●🟥				13	7.00E-04	1.41E-02
36	PTPRD	protein tyrosine phosphatase, receptor type, D		●🟥			17	7.00E-04	1.47E-02
37	LETM1	leucine zipper-EF-hand containing transmembrane protein 1		●🟥			9	7.00E-04	1.47E-02
38	ZMYM4	zinc finger, MYM-type 4	●🟥				12	8.00E-04	1.55E-02
39	CDH12	cadherin 12, type 2 (N-cadherin 2)		●🟥			10	8.00E-04	1.58E-02
40	RP1	retinitis pigmentosa 1 (autosomal dominant)		●🟥			18	1.00E-03	1.87E-02
41	ERBB3	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	●🟥				17	1.50E-03	2.80E-02
42	KIF21B	kinesin family member 21B			●🟩		14	1.80E-03	2.90E-02
43	TGS1	trimethylguanosine synthase homolog (S. cerevisiae)			●🟥		12	2.30E-03	3.48E-02
44	COL19A1	collagen, type XIX, alpha 1		●🟥			12	2.60E-03	4.37E-02
45	WSCD2	WSC domain containing 2	●🟥				12	2.50E-03	4.50E-02
46	PEG3	paternally expressed 3	●🟥				19	2.70E-03	4.54E-02
47	KIAA2022	KIAA2022	●🟥	●🟥			14	2.70E-03	4.54E-02
48	GNAS	GNAS complex locus			●🟥		11	3.60E-03	4.59E-02
49	CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)			●🟥		10	3.50E-03	4.59E-02
50	TBL1XR1	transducin (beta)-like 1 X-linked receptor 1			●🟥		10	3.40E-03	4.59E-02
51	CNTLN	centlein, centrosomal protein		●🟥			13	3.10E-03	4.96E-02