Analysis of gene copy number changes in tumor phylogenetics

Jijun Tang

jtang@cse.sc.edu

Tuesday 4th April, 2017

Overview

1. Background
   - Fluorescence in Situ Hybridization(FISH)
   - Rectilinear Minimum Spanning Tree(RMST)
   - FISHtree(An earlier method)

2. An iterative approach for phylogenetic analysis of tumor progression using FISH copy number(iFISHtree)
   - Methods and experimental design
   - Results

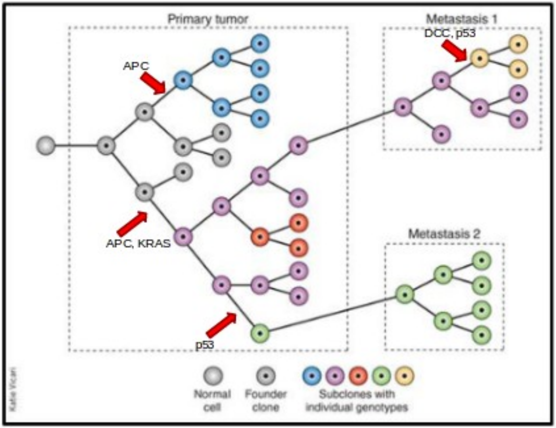
3. Maximum parsimony analysis of gene copy number data(mpFISHtree)
   - Methods and experimental design
   - Results

4. Large scale change(WGD) considered

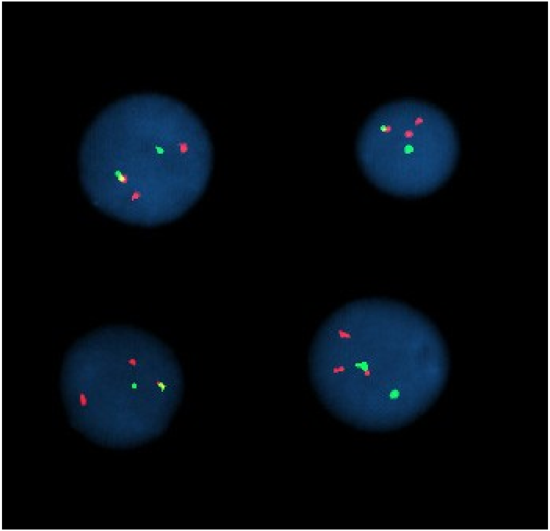
Background

Cancer evolution

Branched Chain Evolution of Cancer



Fluorescence in Situ Hybridization (FISH)



FISH data and distance matrix

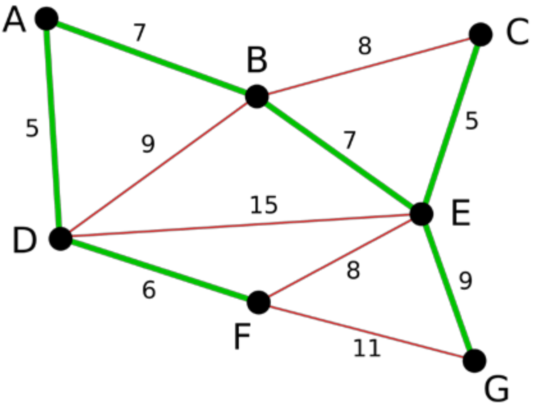
FISH data

	LAMP3	**PROX1**	**PRKAA1**
Cell 1	2	1	2
Cell 2	4	1	3
Cell 3	3	3	2

Distance matrix

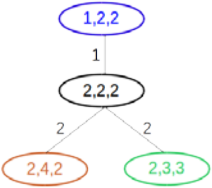
	Cell 1	**Cell 2**	**Cell 2**
Cell 1	0	3	3
Cell 2	3	0	4
Cell 3	3	4	0

Minimum Spanning Tree

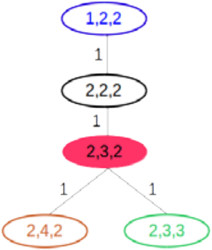


Rectilinear Minimum Spanning Tree

	Gene A Copy #	Gene B Copy #	Gene C Copy #
Count Pattern 1	2	2	2
Count Pattern 2	1	2	2
Count Pattern 3	2	4	2
Count Pattern 4	2	3	3



Minimum Spanning Tree



Rectilinear Steiner Minimum Tree

FISHtree (An earlier method by Chowdhury *et al*)

```
**Input**: a set $S$ of $k$ cell count patterns on $d$ gene probes
**Output**: a tree with additional steiner nodes if needed and $k$ nodes that correspond to $k$ input cell count patterns respectively
**Initialization**: the initial tree $T_0$ = a Minimum Spanning tree on $k$ cell count patterns under the rectilinear metric
Calculate Minimum Spanning Network ($MSN$) on $S$
Identify all 3-node subsets of $MSN$, $T$, where at least two pairs of nodes out of the 3 nodes are connected
**for** each element $T_i$ of $T$ **do**
    Identify candidate Steiner node set $L$ by taking combination of the values of coordinate axes of the points in $T_i$
    **for** each element $L_i$ of $L$ **do**
        Identify $MST$ on $\{S \cup L_i\}$
        Let $current_{m}st_{w}eight = weight(\{S \cup L_i\})$ **if**
          $current\_mst\_weight < min\_weight$ **then**
            $min\_weight = curren\_mst\_weight$
            $S = S \cup L_i$
            $steiner\_tree = MST(\{S\})$
**Output** steiner_tree and min_weight
```

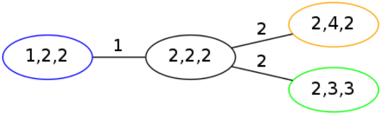
FISHtree → Datasets

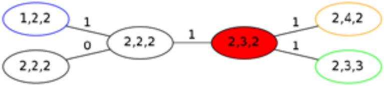
Cancer	Gene marker	Primary	Metastasis
Cervical	LAMP3 PROX1 PRKAA1 CCND1	31	16
Breast	COX-2 MYC CCND1 HER-2 ZNF217 DBC2 CDH1 p53	13	12

Table: Real dataset. The dataset contains cervical and breast cancer samples.

Infer RMST from MST and full binary tree

	Gene A	Gene B	Gene C
Copy Number Profile 1	122		
Copy Number Profile 2	2	2	2
Copy Number Profile 3	242		
Copy Number Profile 3	233		

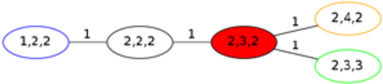




FBT

MST

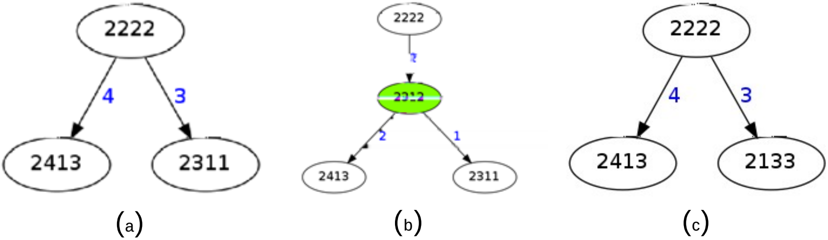
\ (



RMST

An iterative approach for phylogenetic analysis of cancer FISH data(iFISHtree)

iFISHtree → Median idea



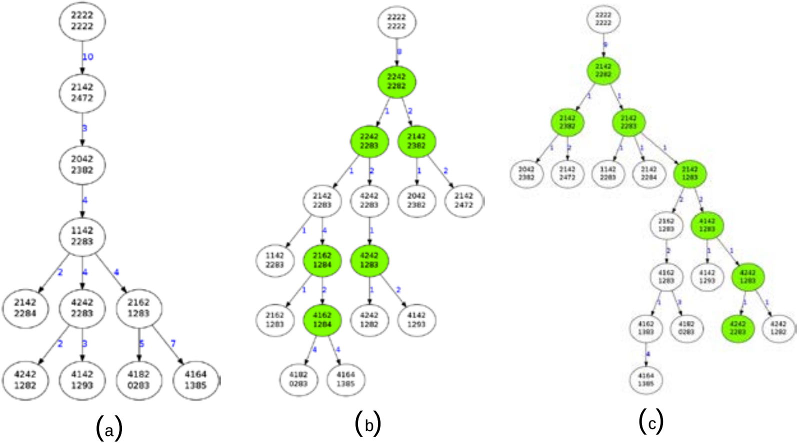
(a)

(b)

(c)

Figure:Instances of RMST(3,d) and the introduction of the steiner node as the median.

iFISHtree → order matters



(a) (b) (c)

Figure:Different orders of adding steiner nodes result in different weights of the resulting trees. (B): 37, (C):36

iFISHtree → inference score

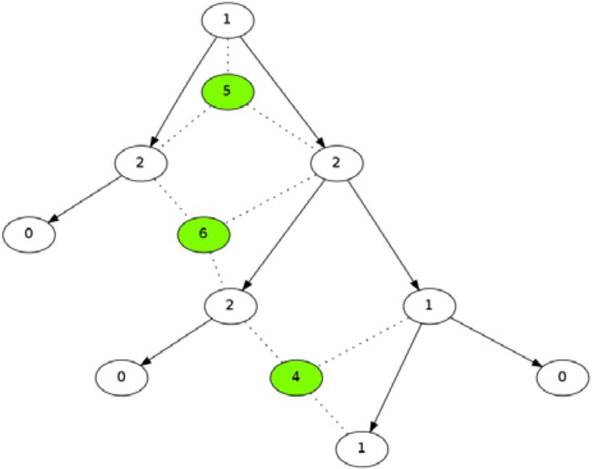


Figure:The definition of *steiner count* of the node in the current tree and the *inference score* of potential steiner nodes to be added.

iFISHtree → Algorithm design

Input: a set of k cell count patterns on d gene probes

Output: a tree with additional steiner nodes if needed and k nodes that correspond to k input cell count patterns respectively

Initialization: the initial tree T_0 = a Minimum Spanning tree on k cell count patterns under the rectilinear metric

Iteration: from tree T_i (V_i) on node set V_i to $T_{i+1}(V_{i+1})$ on node set V_{i+1}

```
    Identify the set $S$ of potential steiner nodes from all possible triplets in $T_i$
    While $S$ is not empty
        Select the potential steiner node $p$ with minimum inference score in $S$
        Build a Minimum Spanning tree on $\{V_i \cup p\}$ as $T(V_i \cup p)$
        If the weight of $T(V_i \cup p)$ is lower than the weight of $T_i(V_i)$
            $T_{i+1}(V_{i+1}) = T(V_i \cup p)$
        Else
            $S = S \setminus \{p\}$
```

Exit condition: S is empty

Breast cancer patient 13 metastasis sample

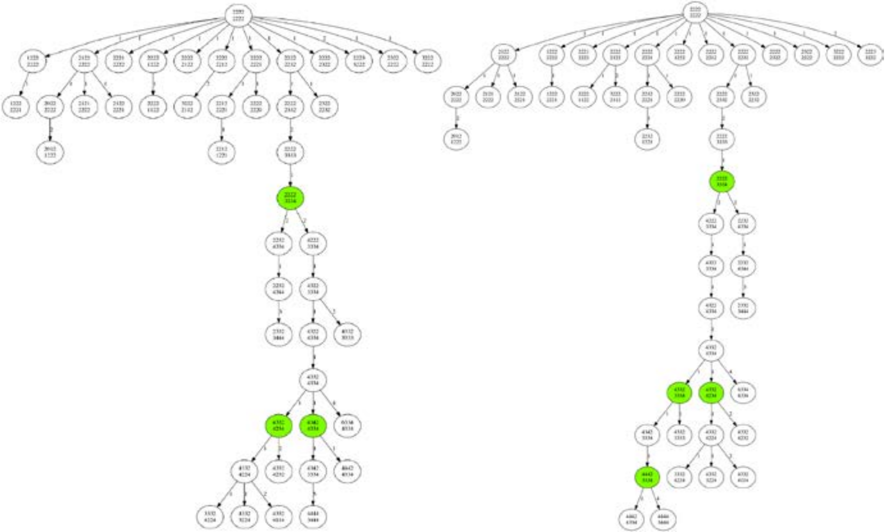


Figure:Score. FISHtree: 87; iFISHtree: 85.

Breast cancer result

Case #	Initial		FISHtrees		iFISHtrees	
	Node #	weight	Node #	weight	Node #	weight
B1_IDC	119	230	135	213	132	**212**
B1_DCIS	143	259	158	**241**	159	242
B2_IDC	104	238	124	217	123	**216**
B3_DCIS	106	72	80	100	80	**98**
B4_IDC	110	232	129	214	129	**213**
B6_IDC	85	116	90	112	90	**111**
B7_IDC	59	128	73	116	71	**113**
B7_DCIS	76	202	84	186	83	**184**
B9_IDC	94	251	121	222	119	**217**
B9_DCIS	76	177	89	164	89	**162**
B10_DCIS	95	154	89	146	89	**145**
B11_DCIS	80	144	87	136	84	**135**
B12_IDC	112	212	124	201	123	**200**
B13_IDC	84	140	92	133	92	**131**
B13_DCIS	43	66	47	63	47	**62**

Table:Comparison on dataset for real breast cancer samples.

Cervical cancer result

Case #	Initial		FISHtrees		iFISHtrees	
	Node #	weight	Node #	weight	Node #	weight
C5	140	208	153	**195**	151	196
C9	130	144	131	143	132	**142**
C10	72	87	72	87	73	**86**
C12	63	72	63	72	64	**71**
C15	66	75	67	74	68	**73**
C21	63	77	67	**73**	65	74
C27	49	60	50	59	52	**57**
C29	76	85	78	83	78	**82**
C32	160	216	167	209	169	**207**
C34	67	88	72	83	73	**82**
C37	71	74	72	73	73	**72**
C42	157	207	164	199	166	**198**
C45	126	183	136	172	140	**169**
C46	87	116	92	110	93	**109**
C49	128	166	132	162	133	**161**
C51	76	83	76	83	83	**76**
C53	64	82	67	82	66	**79**
C54	123	152	129	146	130	**145**

Table: Comparison on dataset for real cervical cancer samples.

Simulation data result

Probe #	Growth factor	FISHtrees =iFISHtrees	FISHtrees >iFISHtrees	FISHtrees <iFISHtrees
4	0.4	176	23	1
6	0.4	161	30	9
8	0.4	162	31	7
4	0.5	182	18	0
6	0.5	160	31	9
8	0.5	152	32	6

Table: Comparison on simulated datasets.

Conclusion

RMST was shown to be a good model for phylogenetic analysis by using FISH cell count pattern data, but it need efficient heuristics because it is a NP-hard problem.

We presented our heuristic method iFISHtree to approximate the RMST based on medium idea.

Our experiments on simulation and real datasets demonstrate the superiority of our algorithm over previous method.

Our method runs at similar and relatively faster speed than earlier method and is supposed to be better with increasing number of gene markers.

Maximum parsimony analysis of gene copy number data

Maximum Parsimony Method(TNT)

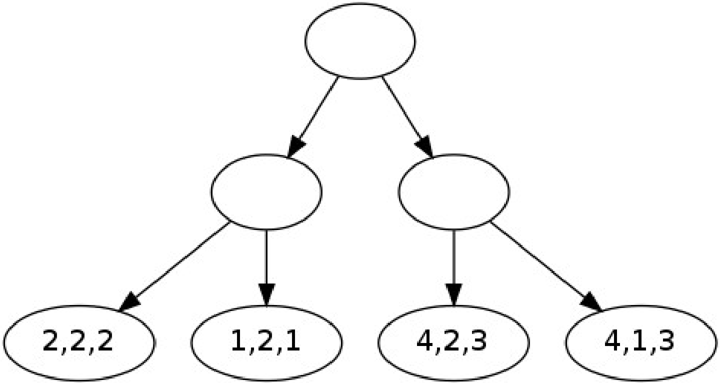


Figure:Tree generated from parsimony phylogeny methods like TNT.

Fitch(bottom up)

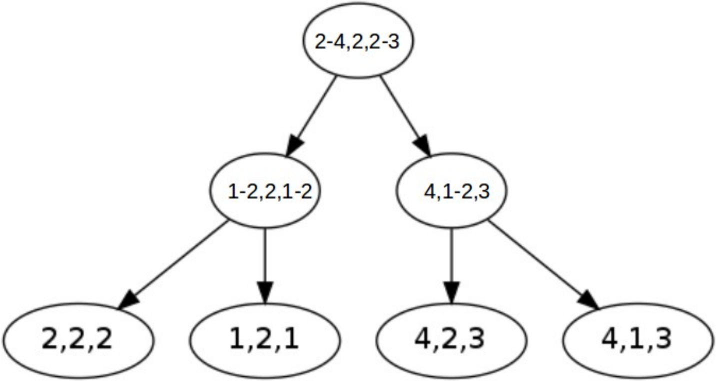


Figure:Fitch algorithm: bottom up.

Fitch(up down)

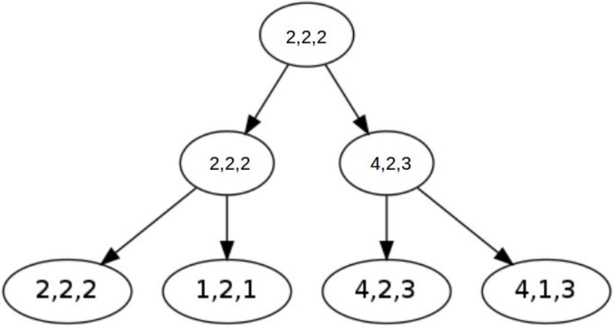
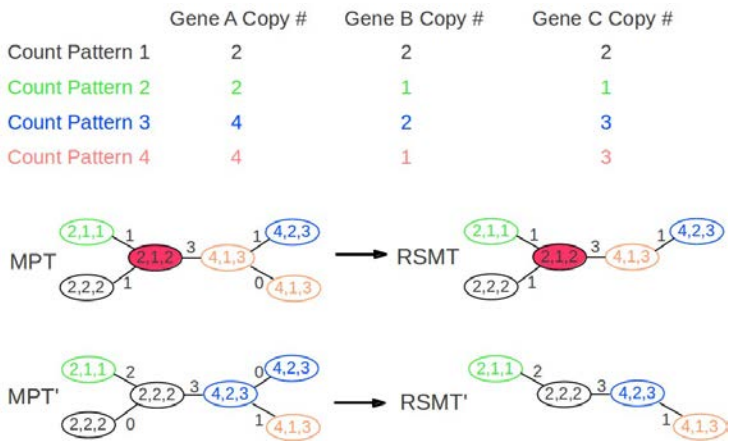


Figure:Fitch algorithm: up down.

MPT →RMST



	Gene A Copy #	Gene B Copy #	Gene C Copy #
Count Pattern 1	2	2	2
Count Pattern 2	2	1	1
Count Pattern 3	4	2	3
Count Pattern 4	4	1	3

Figure:(Top) the input data. (Bottom) two maximum parsimony trees MPT and MPT'. The corresponding RMST and RMST', both of weight 6, shows different steiner nodes number.

Minimizing steiner nodes

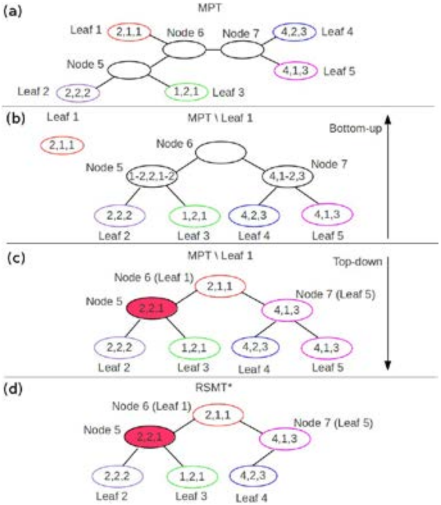


Figure: An example to test whether $Leaf_1$ can be optimally “lifted” to its parent node $Node_6$ in MPT.

Result—FISHtree

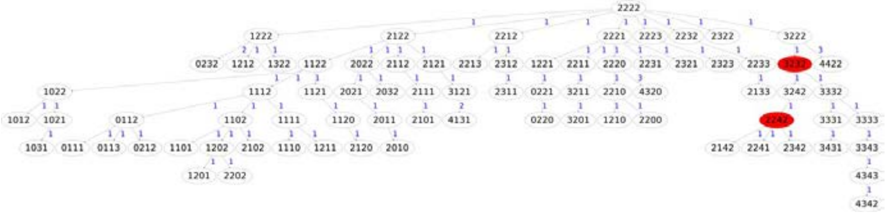


Figure: Given the metastatic cervical cancer sample of patient 12, approximate RMST constructed by FISHtree with weight 83, Each white node represents an input cell count pattern, and each red node represents an inferred Steiner node. Branch lengths are shown in blue.

Result—iFISHtree

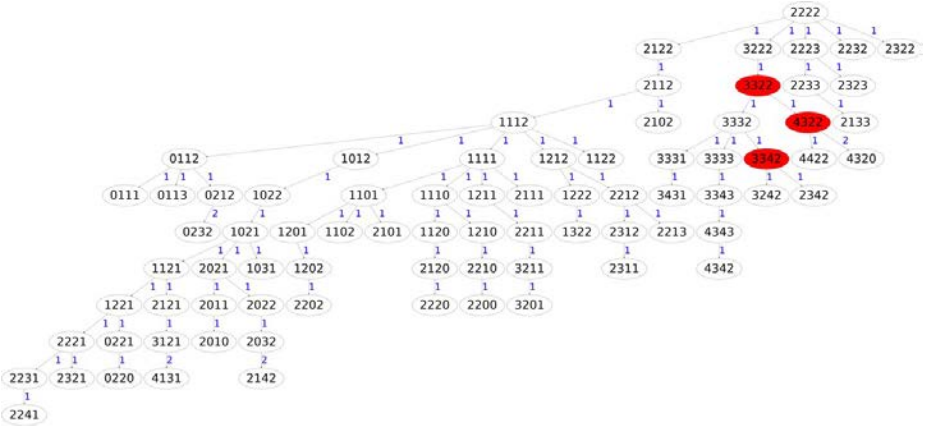


Figure: Given the metastatic cervical cancer sample of patient 12, approximate RMST constructed by iFISHtree with weight 82.

Result—mpFISHtree

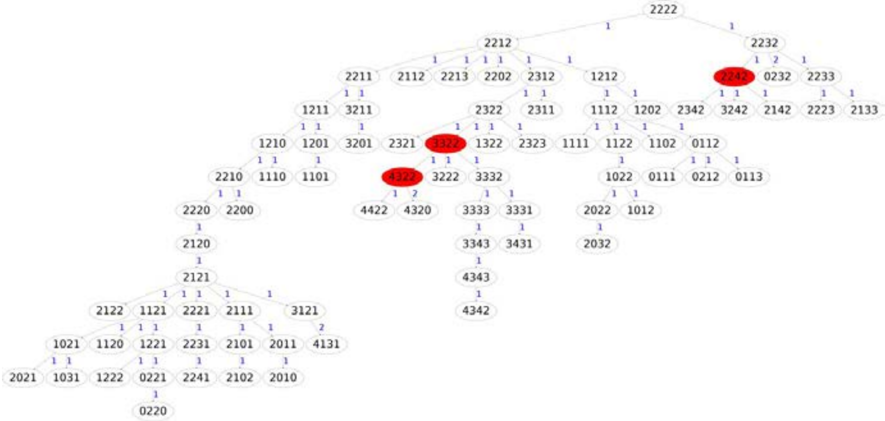


Figure: Given the metastatic cervical cancer sample of patient 12, approximate RMST constructed by mpFISHtree with weight 81.

Breast cancer result

Case #	Tree weight (# Steiner nodes)			
	FISHtree	iFISHtree	mpFISHtree	Exact
B1_IDC	213 (15)	212 (13)	**211** (19)	NA
B1_DCIS	241 (14)	242 (15)	**239** (22)	NA
B2_IDC	217 (15)	216 (20)	**211** (22)	NA
B2_DCIS	56 (2)	56 (2)	**55** (3)	NA
B3_DCIS	100 (7)	**98** (7)	**98** (10)	NA
B4_IDC	214 (16)	**213** (17)	**213** (17)	NA
B6_IDC	112 (4)	**111** (4)	**111** (6)	NA
B7_IDC	116 (8)	**113** (12)	**113** (12)	NA
B7_DCIS	186 (13)	184 (14)	**182** (22)	NA
B9_IDC	222 (22)	217 (25)	**213** (30)	NA
B9_DCIS	164 (12)	163 (13)	**161** (15)	NA
B10_IDC	128 (4)	128 (4)	**127** (4)	NA
B10_DCIS	146 (6)	**145** (8)	**145** (9)	NA
B11_DCIS	136 (6)	135 (7)	**134** (7)	NA
B12_IDC	201 (9)	200 (10)	**198** (15)	NA
B12_DCIS	161 (9)	161 (10)	**158** (13)	NA
B13_IDC	132 (7)	**131** (8)	**131** (8)	NA
B13_DCIS	63 (3)	**62** (4)	**62** (4)	NA

Table: Comparison on dataset for real breast cancer samples.

Cervical cancer result

Case #	Tree weight (# Steiner nodes)			
	FISHtree	iFISHtree	mpFISHtree	Exact
C5	195 (13)	196 (12)	**194** (13)	**194** (13)
C6	82 (2)	82 (2)	**81** (5)	**81** (4)
C8	103 (6)	103 (6)	**100** (9)	**100** (8)
C9	143 (1)	**142** (2)	**142** (5)	**142** (2)
C10	87 (0)	**86** (1)	**86** (1)	**86** (1)
C12	72 (1)	**71** (2)	**71** (2)	**71** (2)
C13	150 (5)	150 (5)	**149** (7)	**149** (7)
C15	74 (1)	**73** (2)	**73** (2)	**73** (2)
C18	127 (4)	127 (4)	**126** (6)	**126** (6)
C21	**73** (4)	74 (3)	**73** (5)	**73** (4)
C27	59 (1)	**57** (3)	**57** (2)	**57** (3)
C29	83 (2)	82 (3)	**81** (3)	**81** (3)
C30	118 (9)	118 (9)	**116** (9)	**116** (10)
C32	209 (7)	207 (9)	**205** (14)	**205** (13)
C34	83 (5)	**82** (6)	**82** (6)	**82** (6)
C35	67 (1)	67 (1)	**66** (2)	**66** (3)
C42	199 (7)	198 (9)	**197** (12)	**197** (11)
C45	172 (10)	**169** (13)	**169** (14)	**169** (15)
C46	110 (5)	109 (6)	**108** (8)	**108** (7)
C49	162 (4)	**161** (5)	**161** (7)	**161** (7)
C53	80 (3)	**79** (4)	**79** (4)	**79** (4)
C54	146 (6)	145 (7)	**144** (10)	**144** (9)

Table:Comparison on dataset for real cervical cancer samples.

Simulation data result

Probe #	Growth factor	Best score count (Best score percentage)			
		FISHtree	iFISHtree	mpFISHtree	Exact
4	0.4	92 (46%)	137 (68.5%)	196 (98%)	200
6	0.4	70 (35%)	98 (49%)	194 (97%)	N/A
8	0.4	41 (20.5%)	69 (34.5%)	196 (98%)	N/A
4	0.5	93 (46.5%)	130 (65%)	194 (97%)	200
6	0.5	68 (34%)	99 (49.5%)	196 (98%)	N/A
8	0.5	40 (20%)	64 (32%)	195 (97.5%)	N/A

Table:Comparison on simulated datasets.

Large scale change(WGD)

WGD exists in 37% of cancer.

Considering large scale change can greatly extend the use of our method.

Chowdhury *et al* have some work in considering large scale gene change.

Find the minimum steiner tree considering large scale change is called Duplication Steiner Minimum Tree (DSMT).

Method

Identify possible large scale changes including WGD.

Remove such branches in the tree generated by Chowdhury *et al*, split the tree into disjoint subtrees.

Reconstruct a new RSMT tree for each subtrees using MPT method.

Re-insert the removed branches and thus assemble the final output DSMT tree.

DSMT–Breast cancer

Cell Line	DSMT Best score	
	FISHtree	MPTtree
B1_IDC	217	**206**
B1_DCIS	150	**140**
B2_IDC	203	**189**
B3_DCIS	99	**97**
B4_IDC	203	**193**
B5_IDC	64	**63**
B6_IDC	108	**106**
B6_DCIS	**42**	43
B7_IDC	116	**115**
B10_IDC	125	**123**
B11_DCIS	122	**121**
B12_IDC	125	**123**
B12_DCIS	162	**149**
B13_IDC	132	**129**
B13_DCIS	63	**61**

Table:Comparison on the real datasets for DSMT on breast cancer samples.

DSMT–Cervical cancer

Cell Line	DSMT Best score	
	FISHtree	MPTtree
C6	82	**81**
C8	95	**93**
C18	126	**122**
C24	**201**	204
C29	80	**76**
C34	**81**	82
C53	75	**71**

Table:Comparison on the real datasets for DSMT on cervical cancer samples.

DSMT–Simulation data

Probe #	Growth factor	DMST Best score count (Best score percentage)	
		FISHtree	MPTtree
4	0.4	175 (87.5%)	191 (95.5%)
6	0.4	145 (35%)	194 (97%)
8	0.4	101 (50.5%)	199 (99.5%)
4	0.5	178 (89%)	189 (94.5%)
6	0.5	147 (73.5%)	193 (96.5%)
8	0.5	93 (46.5%)	200 (100%)

Table: Comparison on simulated datasets for DMST.

Conclusion

We presented our heuristic method MPFISHtree to approximate the RMST based on Maximum Parsimony phylogeny reconstruction (TNT).

We extend our MPFISHtree to consider large genome change including WGD as DMST.

Our experiments on simulation and real datasets demonstrate the superiority of our algorithms over previous methods.

Our method tried to produce the solution with the minimum number of steiner nodes.

Our method can be extended to apply on other data type such as copy number variation(CNV) data.

The End