Supplementary Material.

S1. Elements of a Markov chain stochastic process

The Markov chain dynamical system (*S1, S2, S3*) is written as:
$\vec{v}_{k+1} = \vec{v}_k A \quad (k = 0, 1, 2, \dots),$
where A is a 50 X 50 transition matrix with row entries adding to 1 (a stochastic matrix). The (i, j)th entry, denoted a_{ij} is interpreted as a transition probability for disease progression from site `i' to site `j'. The 50 sites making up the rows and columns of the matrix are all of the primary and metastatic sites listed in the data set (*15*), with site 23 corresponding to `lung' (we number the sites in alphabetical order). The vector $\vec{v}_k$, called the `state-vector', contains the statistical distribution of metastatic tumors at available sites. The entries of the state-vector also sum to 1 and can be interpreted as probabilities of having metastatic disease at each of the sites. The initial state-vector for our model corresponds to primary lung cancer:
$\vec{v}_0 = (0, 0, \dots, 1, 0, 0, \dots),$
with a 1 placed in position 23.

Given a transition matrix A, the state-vector is then iterated:
$\vec{v}_1 = \vec{v}_0 A$
$\vec{v}_2 = \vec{v}_1 A = \vec{v}_0 A^2$
.
.
.
$\vec{v}_k = \vec{v}_0 A^k,$
to produce the statistical distribution of metastatic tumors at each step `k'. For example, for k=2, the entries of $\vec{v}_2$ give the statistical distribution of all two-step pathways starting at the lung and ending at each of the sites in the model. This is the distribution shown in Figure S2. In the limit $k \to \infty$, the state-vector will iterate to the `steady-state' (long-time) distribution $\vec{v}_\infty$
$\vec{v}_\infty = \lim_{k\to\infty} \vec{v}_0 A^k,$
We calculate the entries of A so that this steady-state matches the target lung cancer distribution $\vec{v}_T$ from the data set. We explain this (briefly) below, and refer the reader to (*S2*) for more details.

As a practical matter, it is useful to use the fact that the steady-state vector is a left eigenvector of the transition matrix (*S3,S4*) with eigenvalue 1, hence:
$\vec{v}_\infty (A - I) = 0,$
We then calculate the 2500 entries of A so that $\vec{v}_\infty \approx \vec{v}_T$, up to some numerical threshold. Since this is an underdetermined problem, we need extra information to arrive at a meaningful `lung cancer' transition matrix A, from the many that satisfy the above requirement.
The solution to this problem is carried out iteratively on a sequence of transition matrix `candidates' A_j, $(j = 0, 1, 2, \dots)$, whose entries we adjust until $\left\| \vec{v}_T - \vec{v}_\infty^{\,j} \right\| < \epsilon$, where $\vec{v}_T$ is the target vector for lung cancer, $\vec{v}_\infty^{\,j}$ is the steady-state vector associated with A_j, and ϵ is our numerical threshold which we typically choose $O(10^{-5})$.

We start with a candidate transition matrix A_0, where the lung row (row 23) is chosen to be the target vector for lung cancer, $\vec{v}_T$, and all other rows are taken to be the generic distribution $\vec{v}_g$ over all primary cancer types. Thus, A_0 is a Rank 2 matrix (two linearly independent rows), but does not have $\vec{v}_T$ as its steady-state vector. We then iteratively and systematically produce a sequence of candidate transition matrices A_j, $(j = 0, 1, 2, \ldots)$ by adjusting the entries (using a random choice scheme) of A_j until we produce a candidate (for some j) so that $\|\vec{v}_T - \vec{v}_\infty^j\| < \epsilon$. Details can be found in (*S2*). New data (*S5*) can then be assimilated into the model by changing the target vector $\vec{v}_T$ to a new target vector that reflects the old data along with the newly assimilated data, weighting the two according to the number of entries in each of the two sets. With the new target vector, the algorithm is run again until convergence to a new transition matrix is achieved.

Figure S1 shows convergence of our algorithm (for large enough j) to a `final' lung cancer transition matrix for the baseline model. Typically, the algorithm converges after around $j \sim 50{,}000$ iterations. The solid curve in this figure is an `ensemble' convergence curve compiled as an average from 1000 runs of our algorithm, which then produces an `ensemble' transition matrix whose entries are the sample averages of the 1000 individual transition matrices produced. The error bars show one standard deviation around the sample average. The `residual' (y-axis) in this graph is $\|\vec{v}_T - \vec{v}_\infty^j\|$. The figure also shows another important feature of our method, which is a non-convergence result. Namely, if we constrain our search method in various ways so that (i) no primary re-seeding is allowed (i.e. we force zeros down the lung column); (ii) no metastasis re-seeding is allowed (i.e. we force zeros down the diagonal), the method does not converge. This is shown in Figure S1 as well. Thus, our model indicates that multi-directionality seems to be an essential feature necessary in order to produce the long-time metastatic distributions consistent with our large data set.

References (Supplementary Materials)

S1. P.K. Newton, J. Mason, J. Nieva, K. Bethel, L.A. Bazhenova, P. Kuhn, A stochastic Markov chain model to describe lung cancer growth and metastasis, *PLoS ONE*, **7(4)**, April e34637, (2012).

S2. D. Gamerman, H.F. Lopes, **Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference**, Chapman & Hall/CRC 2nd Edition, (2006).

S3. J.R. Norris, **Markov Chains**, Cambridge University Press, (1997).

S4. L.E. Stenbygaard, J.B. Sorensen, J.E. Olsen, Metastatic pattern in adenocarcinoma of the lung: An autopsy study from a cohort of 137 consecutive patients with complete resection, *Journal of Thoracic and Cardiovascular Surgery*, **110(4)**, Part 1 1130-1135, (1995).

Figure S1. Convergence plot for the lung cancer matrix. Ensemble averaged (1000 trained matrices) convergence plot associated with algorithm to compute the lung cancer transition matrix. Curve marked with squares shows the convergence (ensemble averaged convergence plots) with no constraints to the lung cancer transition matrix used in this study. Curve marked with circles shows non-convergence when the search is constrained so that metastasis re-seeding and primary re-seeding are not allowed. Curve marked with X's shows non-convergence when the search is constrained so that only metastasis re-seeding is not allowed. Curve marked with diamonds shows non-convergence when the search is constrained so that only primary re-seeding is not allowed.

Figure S2. Metastatic tumor distributions after two-steps, compared with steady-state. Metastatic tumor distribution after two-steps of the Markov chain model (solid), starting from primary lung tumor. Blue shows the `target vector' for lung cancer. Red shows the metastatic tumor distribution obtained from the Markov chain dynamical system after two steps. Comparison shows that after two-steps, the ensemble distributions are nearly the same, hence we focus our analysis on the two-step pathway probabilities.

Table S1.

	Top 30 Untreated/Baseline	Top 30 Stage I	Top 30 Stage II	Untreated	Stage I	Stage II
1	*LN (reg) → LN (reg)	*LN (reg) → LN (reg)	*LN (reg) → LN (reg)	0.02819	0.02766	0.02764
2	Adrenal → LN (reg)	Adrenal → LN (reg)	Adrenal → LN (reg)	0.02461	0.02320	0.02408
3	LN (dist) → LN (reg)	LN (dist) → LN (reg)	LN (dist) → LN (reg)	0.02234	0.01937	0.02057
4	LN (reg) → LN (dist)	LN (reg) → LN (dist)	LN (reg) → LN (dist)	0.01860	0.01613	0.01712
5	Adrenal → LN (dist)	Liver → LN (reg)	Liver → LN (reg)	0.01620	0.01550	0.01566
6	Liver → LN (reg)	LN (reg) → Liver	Adrenal → LN (dist)	0.01501	0.01476	0.01491
7	*LN (dist) → LN (dist)	*LN (reg) → Lung	LN (reg) → Liver	0.01468	0.01461	0.01488
8	LN (reg) → Liver	Adrenal → LN (dist)	*LN (reg) → Lung	0.01440	0.01349	0.01385
9	LN (reg) → Adrenal	LN (reg) → Adrenal	LN (reg) → Adrenal	0.01397	0.01283	0.01355
10	Adrenal → Liver	Kidney → LN (reg)	Adrenal → Liver	0.01253	0.01282	0.01292
11	Kidney → LN (reg)	Adrenal → Liver	*LN (dist) → LN (dist)	0.01245	0.01232	0.01271
12	*Adrenal → Adrenal	*Adrenal → Lung	Kidney → LN (reg)	0.01223	0.01209	0.01241
13	*LN (reg) → Lung	*LN (dist) → LN (dist)	*Adrenal → Adrenal	0.01214	0.01125	0.01185
14	LN (dist) → Liver	Bone → LN (reg)	*Adrenal → Lung	0.01130	0.01115	0.01184
15	LN (dist) → Adrenal	*Adrenal → Adrenal	Bone → LN (reg)	0.01101	0.01083	0.01109
16	Bone → LN (reg)	*Lung → LN (reg)	LN (dist) → Liver	0.01100	0.01042	0.01097
17	*Adrenal → Lung	LN (dist) → Liver	*LN (dist) → Lung	0.01042	0.01023	0.01019
18	Liver → LN (dist)	*LN (dist) → Lung	LN (dist) → Adrenal	0.00988	0.01013	0.01003
19	*LN (dist) → Lung	LN (reg) → Bone	Liver → LN (dist)	0.00952	0.00940	0.00970
20	LN (reg) → Bone	Brain → LN (reg)	*Lung → LN (reg)	0.00931	0.00924	0.00965
21	Pleura → LN (reg)	Liver → LN (dist)	Pleura → LN (reg)	0.00886	0.00904	0.00944
22	Pancreas → LN (reg)	LN (dist) → Adrenal	LN (reg) → Bone	0.00873	0.00894	0.00937
23	Kidney → LN (dist)	Pleura → LN (reg)	*Lung → Adrenal	0.00820	0.00886	0.00839
24	Adrenal → Bone	*Lung → Adrenal	*Liver → Liver	0.00811	0.00872	0.00835
25	*Lung → LN (reg)	*Liver → Liver	Adrenal → Bone	0.00789	0.00819	0.00815
26	*Liver → Liver	*Liver → Lung	Pancreas → LN (reg)	0.00758	0.00817	0.00814
27	Liver → Adrenal	Adrenal → Bone	*Liver → Lung	0.00735	0.00787	0.00781
28	LN (dist) → Bone	Pancreas → LN (reg)	Kidney → LN (dist)	0.00734	0.00775	0.00769
29	Bone → LN (dist)	Kidney → LN (dist)	Liver → Adrenal	0.00728	0.00748	0.00759
30	LN (reg) → Kidney	LN (reg) → Kidney	*Lung → LN (dist)	0.00709	0.00740	0.00715

Table S1. Comparative table of top two-step metastatic pathways of all types from Lung. * paths are multi-directional. See Figure 1 for corresponding diagram. Since each path starts from the lung, we show only the 2nd and 3rd site in the two-step pathway. First column lists the two steps out from lung according to the baseline untreated data set (*15*). Columns 2 and 3 list the two steps out from lung according to the assimilated model which incorporates data set (*32*). Columns 4-6 list the two-step probabilities of the corresponding pathways.

Figure S1

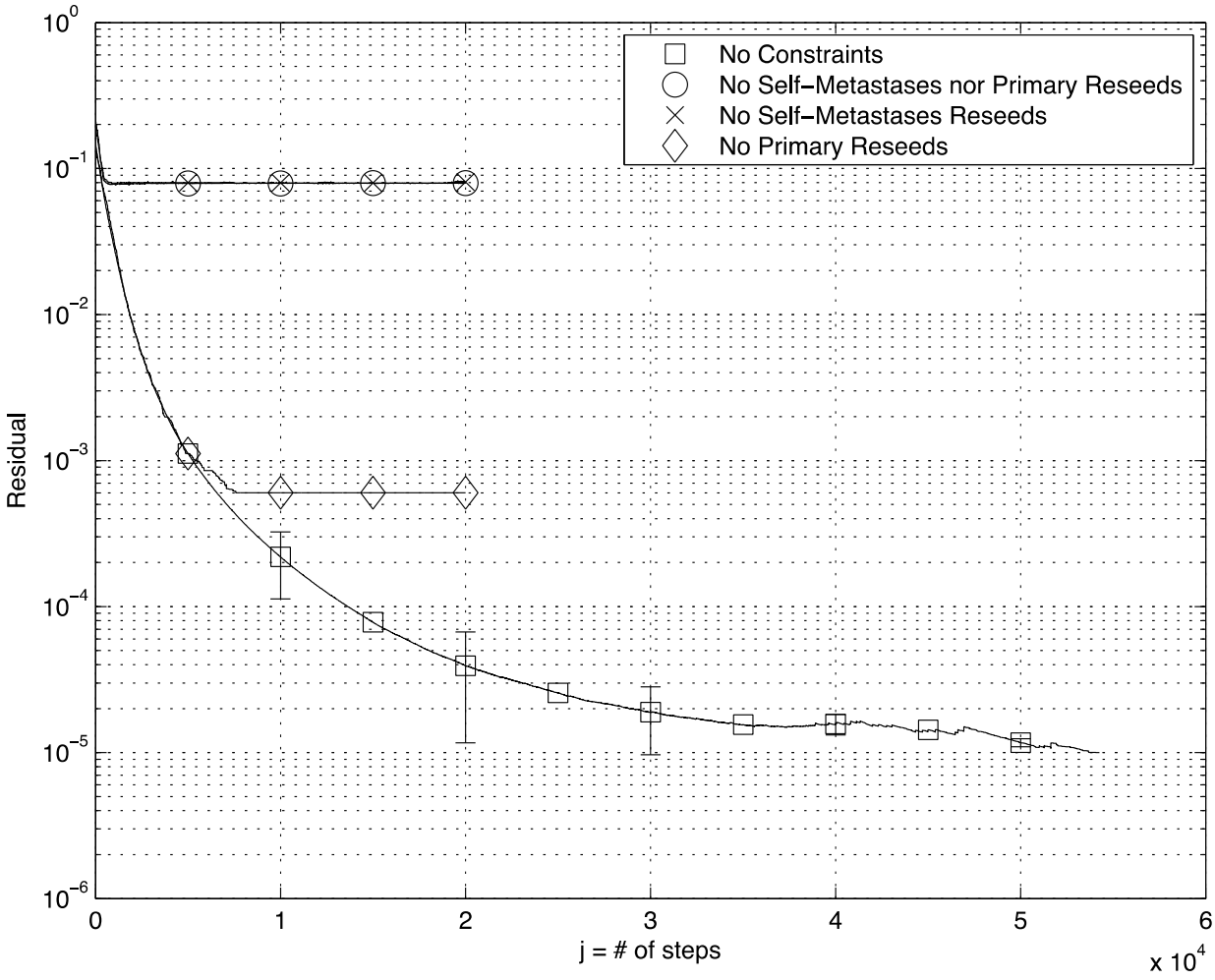
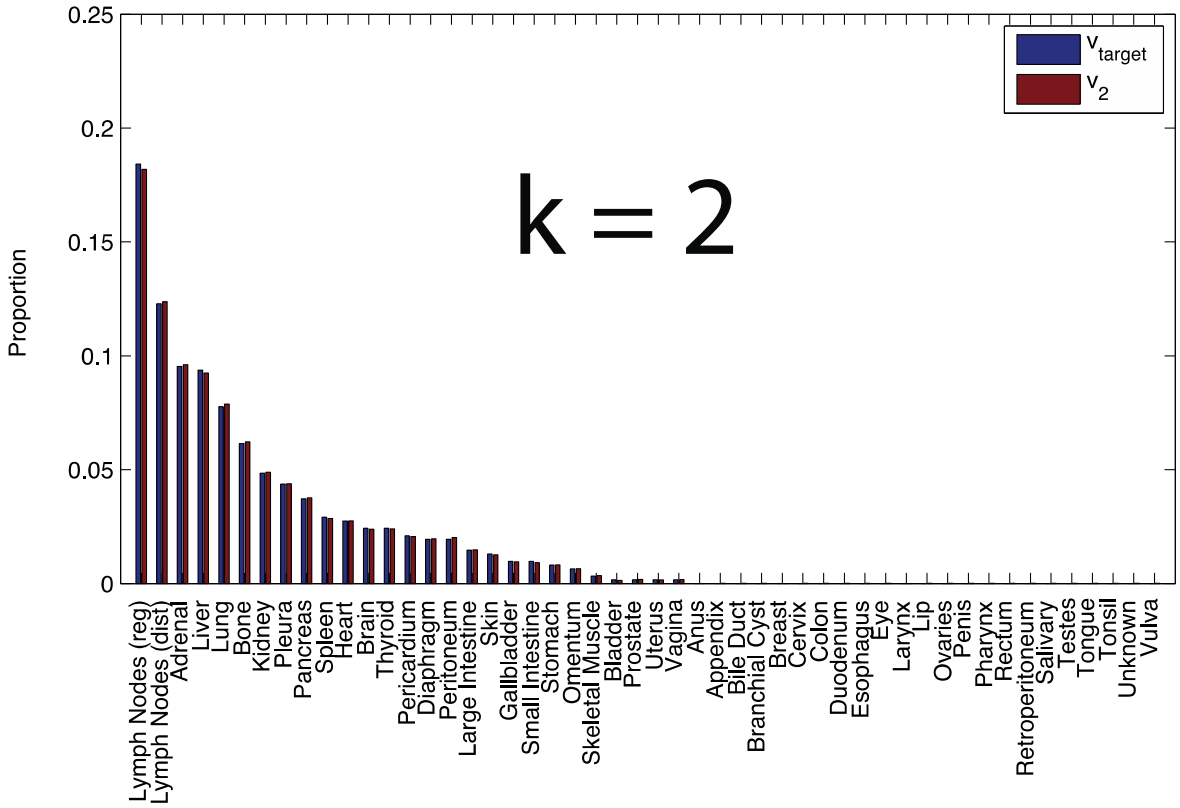

No Constraints
No Self−Metastases nor Primary Reseeds
No Self−Metastases Reseeds
No Primary Reseeds
Residual
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
0
1
2
3
4
5
6
j = # of steps
x 10^{4}

Figure S2


k = 2
Proportion
0.25
0.2
0.15
0.1
0.05
0
v target
v 2
Lymph Nodes (reg)
Lymph Nodes (dist)
Adrenal
Liver
Lung
Bone
Kidney
Pleura
Pancreas
Spleen
Heart
Brain
Thyroid
Pericardium
Diaphragm
Peritoneum
Large Intestine
Skin
Gallbladder
Small Intestine
Stomach
Omentum
Skeletal Muscle
Bladder
Prostate
Uterus
Vagina
Anus
Appendix
Bile Duct
Branchial Cyst
Breast
Cervix
Colon
Duodenum
Esophagus
Eye
Larynx
Lip
Ovaries
Penis
Pharynx
Rectum
Retroperitoneum
Salivary
Testes
Tongue
Tonsil
Unknown
Vulva