Chapter 3
Benchmarking a Simple Yet Effective Approach for Inferring Gene Regulatory Networks from Systems Genetics Data

Sandra Heise, Robert J. Flassig and Steffen Klamt

Abstract We apply our recently proposed gene regulatory network (GRN) reconstruction framework for genetical genomics data to the StatSeq data. This method uses, in a first step, simple genotype–phenotype and phenotype–phenotype correlation measures to construct an initial GRN. This graph contains a high number of false positive edges that are reduced by (i) identifying eQTLs and by retaining only one candidate edge per eQTL, and (ii) by removing edges reflecting indirect effects by means of TRANSWESD, a transitive reduction approach. We discuss the general performance of our framework on the StatSeq in silico dataset by investigating the sensitivity of the two required threshold parameters and by analyzing the impact of certain network features (size, marker distance, and biological variance) on the reconstruction performance. Using selected examples, we also illustrate prominent sources of reconstruction errors. As expected, best results are obtained with large number of samples and larger marker distances. A less intuitive result is that significant (but not too large) biological variance can increase the reconstruction quality. Furthermore, a somewhat surprising finding was that the best performance (in terms of AUPR) could be found for networks of medium size (1,000 nodes), which we had expected to see for networks of small size (100 nodes).

3.1 Introduction

Systems Genetics approaches provide a new paradigm of large-scale genome and network analysis (Jansen and Nap 2001; Jansen 2003; Rockman and Kruglyak 2006; Rockman 2008). These methods use naturally occurring multifactorial perturbations (e.g., polymorphisms) to causally link genetic or chromosomal regions to observed

S. Heise · R. J. Flassig · S. Klamt(✉)
Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstrasse 1, D-39106 Magdeburg, Germany
e-mail: klamt@mpi-magdeburg.mpg.de

A. de la Fuente (ed.), *Gene Network Inference*,
DOI: 10.1007/978-3-642-45161-4_3,

phenotypic trait data. Identifying a chromosomal region (the quantitative trait locus (QTL)) that influences a certain phenotypic trait is known as QTL mapping. In genetical genomics, a particular subclass of systems genetics, gene-expression levels are considered as phenotypic traits (called etraits) and identified QTLs are referred to as expression-QTLs (eQTLs). One application of eQTL maps obtained from genetical genomics approaches is the reconstruction of gene regulatory networks (GRNs).

According to Liu et al. (2010), a GRN reconstruction pipeline for genetical genomics data consists of three major steps: (i) eQTL mapping, (ii) candidate regulator selection, and (iii) network refinement. Step (i) is used to identify chromosomal regions (eQTLs) that impact on expression levels (=traits) of genes. A detailed review on eQTL mapping is, for instance, given by Michaelson et al. (2009). In step (ii), the eQTL map in combination with a genetic map is used to select single candidate (regulator) genes from the eQTLs. Frequently used methods include conditional correlation (Bing and Hoeschele 2005; Keurentjes et al. 2007), local regression (Liu et al. 2008), or analysis of between-strains SNPs (Li et al. 2005). In the third step (iii), network refinement methods are employed to the topology obtained in step (ii), e.g., with the goal to identify and eliminate (false positive) edges arising from indirect effects. Here, Bayesian network approaches (Zhu et al. 2007) and structural equation modeling, SEM, (Liu et al. 2008) have been used.

In this chapter, we apply our recently proposed GRN reconstruction framework for genetical genomics data (Flassig et al. 2013), which incorporates the three major reconstruction steps mentioned above in a modular fashion. The framework follows a *simple-yet-effective* paradigm: it is based on simple correlation measures, without the need for computational demanding optimization steps. This approach is therefore suited for small- and large-scale networks and performed comparable well in the case of little samples but many genes, as we illustrate in Flassig et al. (2013) using simulated and biological data. The workflow of the framework is shown in Fig. 3.1. The initial GRN is constructed based on genotype–phenotype and phenotype–phenotype correlation analyses. Due to genetic linkage there are often groups of genetically adjacent regulator gene candidates, which target the same gene resulting into eQTLs. To avoid many false-positive interaction predictions, single candidate regulators are therefore identified from the eQTLs. Finally, as a method for network refinement in step (iii), indirect path effects are removed by TRANSWESD, a transitive reduction approach introduced recently (Klamt et al. 2010).

3.2 Methods

Figure 3.1 shows the general workflow of our reconstruction framework together with a simple illustrative example. Starting from a typical set of genetical genomics data that include genotyped markers, phenotyped genes and gene-to-marker association, marker linkage analysis, and genotype assignment for each gene is performed in a preprocessing step. In particular, a linkage map is generated in which two markers are indicated to be genetically linked if their genotype–genotype correlation

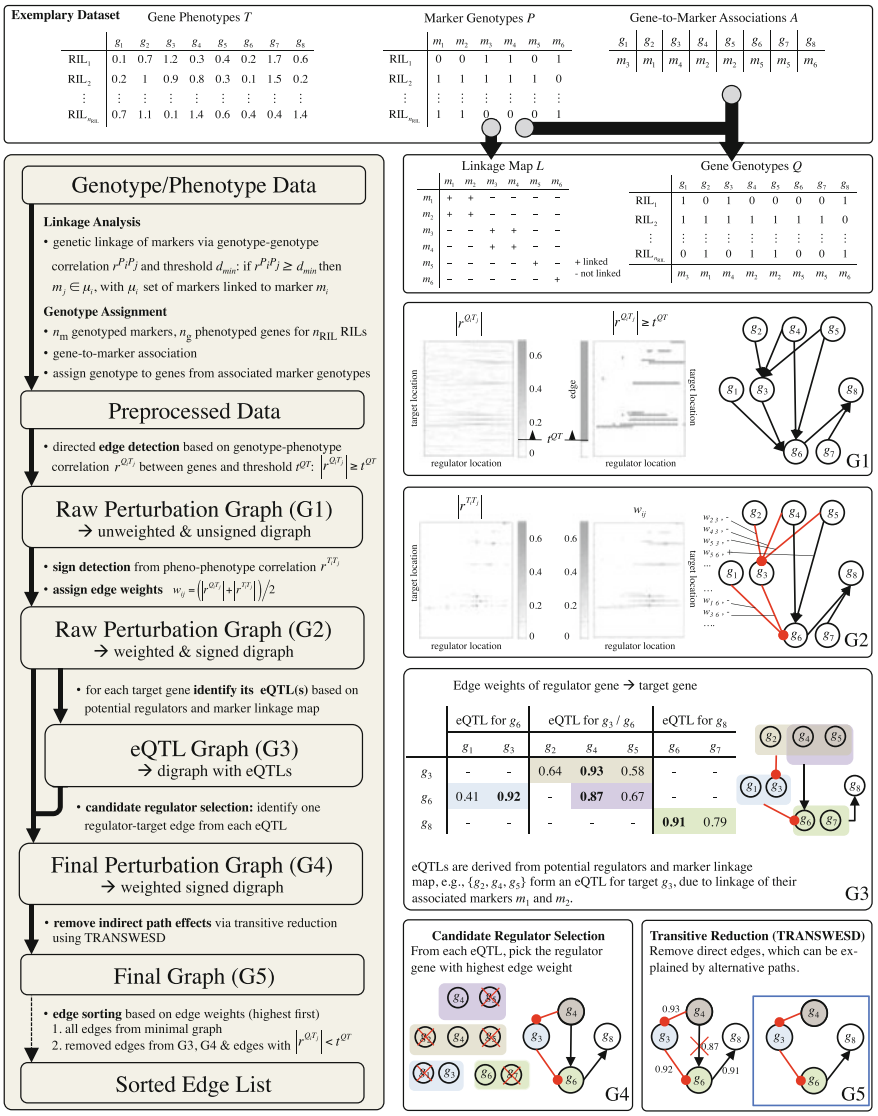


Fig. 3.1 Workflow of the proposed framework for reconstructing GRNs from genetical genomics data (*left*) with an illustrative example (*top panel* and *right*). For detailed explanations see text. Reproduced with permission of Oxford University Press from Flassig et al. (2013)

exceeds a given threshold parameter $d_{\min}$. Then, in a first step, an unweighted and unsigned perturbation graph G1 is derived in which an edge $i \rightarrow j$ is included if their corresponding genotype-phenotype correlation exceeds a second threshold t^{QT}. The nodes in the graph directly correspond to genes while the linkage map (of

the markers) is kept to allow later eQTL assignment for each gene. The perturbation graph G1 is refined to G2 by quantifying each identified edge with respect to edge sign and weight, which indicate activation/repression and interaction strength, respectively. Due to genetic linkage true regulators may be masked by other genes (e.g., on adjacent positions on the genetic map) resulting into eQTLs. The eQTLs of a given target gene t are identified on the basis of all potential regulator genes of t (contained in G2) together with the marker linkage map. These relationships are captured in graph G3, which is the only graph where the nodes represent eQTLs. Graph G4 is subsequently obtained by selecting one candidate regulator per eQTL based on the maximum of the edge weights. We call G4 the final perturbation graph, whose edges reflect direct and indirect effects between genes induced by genetic variations. To identify and remove indirect edges in G4 that can be explained by the operation of sequences of edges (paths) we apply the transitive reduction method TRANSWESD (TRANSitive reduction in WEighted Signed Digraphs) resulting in the final graph G5 containing the identified gene interactions. Optionally, if one is left to verify the interactions experimentally, it is desirable to have a list of edges sorted with respect to edge confidences. Such a list is also required by the evaluation procedure of the StatSeq Systems Genetics Benchmark to assess the quality of a reconstructed network (Sect. 3.3). We generate such a sorted list based on the edge weights. More details on the framework can be found in Flassig et al. (2013).

3.3 Application to the StatSeq Systems Genetics Benchmark: Results and Discussion

We applied our reconstruction framework described in Sect. 3.2 to the in silico StatSeq dataset provided to all contributors of this book. In this section, we will discuss the general performance of the algorithm and investigate the impact of certain network features (size, marker distance, and biological variance) on the reconstruction performance of our applied reconstruction framework. Using selected examples, we will also illustrate prominent sources of reconstruction errors (Sect. 3.3.2).

3.3.1 General Performance Analysis with Respect to Network Configurations

Table 3.1 shows the AUPR and AUROC reconstruction performance (obtained by using optimal values for the thresholds $d_{\min}$ and t^{QT}) for all studied 72 network configurations: 3 different network sizes (100, 1000, 5000) $\times$ 3 replicates (with same topological parameters) $\times$ 2 marker distances (close and far) $\times$ 2 different biological variances (high and low) $\times$ 2 different population sizes (300 and 900) (see also Chap. 1). The performance measures are given for graph G2, G4, and G5

Table 3.1 Reconstruction performance obtained for each network configuration achieved with the indicated optimal parameter values

Configuration	t^{QT}	$d_{\min}$	G2				G4				G5			
			AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP
[100.1.1]: F / L / 300	0.26	0.5	0.7810	0.2247	104	516	0.7842	0.2782	69	38	0.7843	0.2801	69	36
[100.1.2]: F / L / 900	0.22	0.6	0.8058	0.2007	129	669	0.8087	0.2772	81	64	0.8082	0.2750	79	57
[100.1.3]: F / H / 300	0.14	0.2	0.8011	0.2178	97	771	0.8143	0.2527	67	125	0.8145	0.2548	67	118
[100.1.4]: F / H / 900	0.1	0.5	0.8370	0.3041	125	834	0.8491	0.3588	92	84	0.8494	0.3635	92	72
[100.1.5]: C / L / 300	0.6	1	0.7553	0.1043	27	111	0.7552	0.1216	25	83	0.7552	0.1219	25	81
[100.1.6]: C / L / 900	0.24	0.1	0.7561	0.0942	126	1706	0.7668	0.1253	35	61	0.7666	0.1245	34	59
[100.1.7]: C / H / 300	0.16	0.3	0.7792	0.1977	120	1596	0.8034	0.2593	62	69	0.8036	0.2611	62	65
[100.1.8]: C / H / 900	0.14	0.2	0.8062	0.2197	119	1529	0.8362	0.3104	67	41	0.8363	0.3111	67	39
[100.2.1]: F / L / 300	0.2	0.2	0.7675	0.2467	139	778	0.7644	0.2898	85	69	0.7647	0.3016	82	54
[100.2.2]: F / L / 900	0.16	0.5	0.7858	0.2050	179	1202	0.7722	0.2910	98	86	0.7727	0.3208	97	59
[100.2.3]: F / H / 300	0.12	0.2	0.7626	0.2198	110	1017	0.7781	0.2481	78	227	0.7791	0.2708	77	183
[100.2.4]: F / H / 900	0.12	0.5	0.8043	0.2972	124	563	0.8104	0.3432	96	58	0.8106	0.3613	92	43
[100.2.5]: C / L / 300	0.22	0.2	0.7307	0.1529	146	1821	0.7257	0.1902	54	81	0.7256	0.1914	53	71
[100.2.6]: C / L / 900	0.12	0.7	0.7355	0.1333	214	3386	0.7134	0.1831	61	142	0.7142	0.1939	61	115
[100.2.7]: C / H / 300	0.02	0.2	0.7524	0.2005	285	7953	0.7624	0.2430	92	322	0.7617	0.2379	80	220
[100.2.8]: C / H / 900	0.1	0.6	0.7686	0.2144	152	2104	0.7987	0.2965	65	71	0.7988	0.3038	65	64
[100.3.1]: F / L / 300	0.24	0.2	0.7725	0.2324	92	508	0.7755	0.2737	68	42	0.7756	0.2772	68	35
[100.3.2]: F / L / 900	0.2	0.4	0.7834	0.2505	127	821	0.7897	0.2886	84	68	0.7898	0.3013	83	49
[100.3.3]: F / H / 300	0.12	0.3	0.7456	0.1914	105	948	0.7596	0.2266	83	250	0.7613	0.2319	82	187
[100.3.4]: F / H / 900	0.14	0.6	0.7799	0.2479	86	430	0.7878	0.2901	74	74	0.7864	0.2884	73	72
[100.3.5]: C / L / 300	0.28	0.2	0.7124	0.1084	102	1043	0.7169	0.1566	46	44	0.7169	0.1579	46	43
[100.3.6]: C / L / 900	0.14	0.7	0.7156	0.1108	184	3068	0.7233	0.1676	65	135	0.7202	0.1642	60	119
[100.3.7]: C / H / 300	0.22	0.7	0.7400	0.1188	39	522	0.7533	0.1420	23	28	0.7533	0.1420	23	28
[100.3.8]: C / H / 900	0.24	0.2	0.7359	0.1190	51	463	0.7453	0.1612	27	11	0.7453	0.1612	27	11
Configuration average for 300 samples			0.7584	0.1846	114	1465	0.7661	0.2235	63	115	0.7663	0.2274	61	93
Configuration averaged for 900 samples			0.7762	0.1997	135	1398	0.7835	0.2577	70	75	0.7832	0.2641	69	63

(continued)

Table 3.1 (continued)

Configuration	t^{QT}	$d_{\min}$	G2				G4				G5			
			AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP
[1000.1.1]: F / L / 300	0.15	0.4	0.8273	0.1654	1464	26333	0.8224	0.2037	1115	8658	0.8203	0.2060	998	4523
[1000.1.2]: F / L / 900	0.1	0.3	0.8725	0.2170	1887	28841	0.8647	0.3015	1397	4666	0.8612	0.3171	1275	2938
[1000.1.3]: F / H / 300	0.1	0.2	0.8354	0.1996	1309	91968	0.8499	0.2143	1082	33130	0.8567	0.2230	948	9124
[1000.1.4]: F / H / 900	0.1	0.2	0.8918	0.2842	1127	11114	0.8939	0.3417	1061	2756	0.8938	0.3430	1048	2531
[1000.1.5]: C / L / 300	0.15	0.5	0.8162	0.0872	1579	71242	0.8133	0.1479	748	6271	0.8130	0.1505	694	3588
[1000.1.6]: C / L / 900	0.1	0.2	0.8429	0.0992	2042	106265	0.8243	0.1774	985	5282	0.8218	0.1857	875	2716
[1000.1.7]: C / H / 300	0.1	0.3	0.8136	0.1431	1366	123310	0.8363	0.1773	892	16739	0.8395	0.1846	816	6011
[1000.1.8]: C / H / 900	0.1	0.2	0.8853	0.2094	1239	34817	0.8928	0.3185	1004	2351	0.8927	0.3226	998	2083
[1000.2.1]: F / L / 300	0.15	0.3	0.8676	0.2738	1513	24947	0.8670	0.3463	1241	8061	0.8667	0.3529	1171	4415
[1000.2.2]: F / L / 900	0.1	0.4	0.8976	0.3050	1844	29917	0.8969	0.3801	1513	4978	0.8952	0.4199	1420	2981
[1000.2.3]: F / H / 300	0.1	0.5	0.8554	0.2231	1221	92487	0.8679	0.2265	1076	46949	0.8804	0.2404	970	9557
[1000.2.4]: F / H / 900	0.1	0.3	0.9166	0.3379	1166	12491	0.9193	0.3743	1094	2960	0.9192	0.3854	1085	2658
[1000.2.5]: C / L / 300	0.15	0.4	0.8392	0.1114	1483	69014	0.8407	0.1747	813	5852	0.8407	0.1805	758	3342
[1000.2.6]: C / L / 900	0.15	0.4	0.8763	0.1541	1446	52275	0.8815	0.2718	964	1506	0.8813	0.2801	949	1180
[1000.2.7]: C / H / 300	0.1	0.3	0.8440	0.1592	1300	121251	0.8706	0.1914	900	16997	0.8743	0.2006	847	6211
[1000.2.8]: C / H / 900	0.1	0.2	0.8994	0.2107	1205	40136	0.9084	0.2993	950	2462	0.9085	0.3117	947	2160
[1000.3.1]: F / L / 300	0.2	0.5	0.8566	0.2079	1089	9353	0.8571	0.2716	966	1437	0.8571	0.2808	955	1210
[1000.3.2]: F / L / 900	0.15	0.4	0.8834	0.2195	1451	14163	0.8828	0.2871	1121	1372	0.8828	0.3109	1107	874
[1000.3.3]: F / H / 300	0.1	0.4	0.8398	0.2135	1366	94845	0.8515	0.2224	1183	41839	0.8605	0.2337	1047	9572
[1000.3.4]: F / H / 900	0.1	0.3	0.8948	0.2602	1144	11561	0.8969	0.2977	1054	2818	0.8969	0.3059	1048	2645
[1000.3.5]: C / L / 300	0.2	0.4	0.8246	0.0789	1241	38766	0.8240	0.1001	517	1881	0.8240	0.1017	502	1684
[1000.3.6]: C / L / 900	0.1	0.2	0.8613	0.0935	1941	95880	0.8574	0.1246	807	4475	0.8567	0.1293	728	2707
[1000.3.7]: C / H / 300	0.1	0.3	0.8243	0.1315	1338	122280	0.8481	0.1588	902	17016	0.8515	0.1680	824	6328
[1000.3.8]: C / H / 900	0.1	0.3	0.8844	0.1617	1232	38206	0.8927	0.2197	906	2407	0.8927	0.2252	899	2231
Configuration averaged for 300 samples			0.8370	0.1662	1356	73816	0.8457	0.2029	953	17069	0.8487	0.2102	878	5464
Configuration averaged for 900 samples			0.8839	0.2127	1477	39639	0.8843	0.2828	1071	3169	0.8836	0.2947	1032	2309

(continued)

Table 3.1 (continued)

Configuration	t^{QT}	$d_{\min}$	G2				G4				G5			
			AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP	AUROC	AUPvR	TP	FP
[5000.1.1]: F / L / 300	0.25	0.6	0.8630	0.1839	4470	29115	0.8628	0.2226	934	4287	0.8628	0.2145	931	3781
[5000.1.2]: F / L / 900	0.15	0.4	0.9117	0.2672	6990	71653	0.9114	0.3865	2269	4802	0.9113	0.4022	2247	2688
[5000.1.3]: F / H / 300	0.2	0.2	0.8746	0.1451	2196	20255	0.8747	0.1480	192	8890	0.8747	0.1479	192	8863
[5000.1.4]: F / H / 900	0.1	0.3	0.9210	0.2665	5790	112803	0.9216	0.3004	1550	50216	0.9216	0.3030	1539	46881
[5000.1.5]: C / L / 300	0.2	0.3	0.8652	0.0570	5770	267008	0.8640	0.0910	1115	13438	0.8640	0.0924	1106	12350
[5000.1.6]: C / L / 900	0.15	0.4	0.8976	0.0684	7173	364594	0.8962	0.1266	1475	7823	0.8962	0.1301	1468	6825
[5000.1.7]: C / H / 300	0.2	0.2	0.8715	0.1035	2363	71630	0.8722	0.1319	177	7098	0.8722	0.1317	177	7087
[5000.1.8]: C / H / 900	0.1	0.2	0.9204	0.1584	5871	326990	0.9232	0.2634	1370	32291	0.9232	0.2650	1351	29118
[5000.2.1]: F / L / 300	0.25	0.7	0.7919	0.0708	4782	37381	0.7912	0.0712	685	29707	0.7884	0.0555	649	16368
[5000.2.2]: F / L / 900	0.15	0.1	0.8536	0.1156	7548	87901	0.8428	0.2746	1485	5575	0.8426	0.2903	1478	4050
[5000.2.3]: F / H / 300	0.2	0.3	0.7946	0.0759	2695	22128	0.7946	0.0798	147	9926	0.7946	0.0798	147	9802
[5000.2.4]: F / H / 900	0.15	0.3	0.8354	0.1223	3964	24811	0.8352	0.2295	313	996	0.8352	0.2300	313	929
[5000.2.5]: C / L / 300	0.2	0.4	0.8019	0.0203	6203	268571	0.7862	0.0319	692	13978	0.7861	0.0321	690	13342
[5000.2.6]: C / L / 900	0.15	0.4	0.8365	0.0257	7476	358538	0.8158	0.0569	1038	8850	0.8157	0.0586	1037	7891
[5000.2.7]: C / H / 300	0.2	0.3	0.7985	0.0434	2981	87816	0.7986	0.0748	143	8600	0.7986	0.0748	143	8486
[5000.2.8]: C / H / 900	0.15	0.3	0.8376	0.0623	4045	113066	0.8375	0.1869	297	2259	0.8375	0.1868	297	2216
[5000.3.1]: F / L / 300	0.2	0.4	0.8677	0.1854	5177	56915	0.8676	0.2282	1404	15288	0.8676	0.2376	1389	13763
[5000.3.2]: F / L / 900	0.1	0.6	0.9143	0.2240	8975	209561	0.9136	0.3275	3181	80936	0.9125	0.3331	2792	50375
[5000.3.3]: F / H / 300	0.15	0.2	0.8663	0.1465	3599	233697	0.8675	0.1579	730	139252	0.8687	0.1598	694	62094
[5000.3.4]: F / H / 900	0.1	0.3	0.9269	0.2571	5508	110540	0.9276	0.2904	1459	52057	0.9276	0.2994	1458	49556
[5000.3.5]: C / L / 300	0.2	0.4	0.8547	0.0707	5288	255626	0.8552	0.1145	940	14745	0.8551	0.1182	938	13384
[5000.3.6]: C / L / 900	0.15	0.3	0.9019	0.0859	6879	320222	0.9024	0.1603	1550	7609	0.9024	0.1665	1537	6399
[5000.3.7]: C / H / 300	0.15	0.3	0.8687	0.0997	3440	341182	0.8723	0.1255	574	76928	0.8729	0.1280	553	42888
[5000.3.8]: C / H / 900	0.1	0.4	0.9200	0.1435	5617	321612	0.9231	0.1921	1341	35060	0.9232	0.1997	1337	32251
Configuration averaged for 300 samples			0.8432	0.1002	4080	140944	0.8422	0.1231	644	28511	0.8421	0.1227	634	17684
Configuration averaged for 900 samples			0.8897	0.1497	6320	201858	0.8875	0.2329	1444	24040	0.8874	0.2387	1405	19932

Each network configuration is described by a network ID [nodes.replicate.configuration] specifying the number of *nodes* (100/1,000/5,000), the *replicate* (1/2/3) and the 8 *configurations*. The 8 configurations 1, . . . , 8 correspond to the IDs in Chap. 1 and depend on marker distance (**C**lose N(1,0.1) / **F**ar N(5,1)) / biological variance (**L**ow N(1,0.1) / **H**igh N(1,0.25)) / population size (300/900). *TP*, True positive; *FP*, false positive. See also Fig. 3.3

to be able to assess the overall effects of the two major pruning steps within our approach (G2 → G4: selection of one candidate edge per eQTL; G4 → G5: removal of edges that most likely stem from indirect effects (TRANSWESD); see Fig. 3.1). We will mainly focus on the AUPR measure since this is the most appropriate one for sparse networks.

As a general trend, we observe that the first (eQTL) pruning step leads in all cases to an improvement of the AUPR, particularly pronounced in the case of large population sizes (see also averaged values in Table 3.1). The second (TRANSWESD) pruning step achieves a significant (but compared to the eQTL pruning lower) AUPR improvement when using the larger population size, whereas only a minor or even no effect can be seen for reconstruction based on the small population with 300 individuals. The effects of the two pruning steps are also well reflected by the number of true positive (TP) and false positive (FP) edges in Table 3.1.

As expected, we see that a larger population size always helps to yield a better reconstruction quality (see also Fig. 3.3). Somewhat surprising was the finding that the best (averaged) AUPR value could be found for the G5 graph of medium size networks (1,000 nodes), here we had expected to see this for networks with 100 nodes.

In the following we will discuss the sensitivity of the reconstruction results with respect to the threshold parameters (t^{Qt} and $d_{\min}$) and the impact of marker distance, biological variance, and population sizes by the example of the first 100-nodes network (networks 100.1.1–100.1.8 in Table 3.1). Similar results can be found for the replicates (100.2.x and 100.3.x) and/or networks of larger size (1000.x.x; 5000.x.x). Figure 3.2 shows for configurations 100.1.1–100.1.8 the resulting AUPR and AUROC performances of the reconstructed G5 networks in the two-dimensional space of meaningful threshold parameters. Clearly, as already outlined above, larger population size (900 samples instead of 300) improves the reconstruction quality (compare odd vs. even numbers of network configurations) although, in line with our results in Flassig et al. (2013), the differences are only moderate. We also see that the optimal threshold regions are similar for all 8 networks. However, one can observe that in the case of low sample size (300) the optimal AUROC/AUPR region is more confined. Thus, the method seems to be fairly robust against a variation of thresholds but an appropriate threshold selection strategy is important for small sample sizes. Generally, the genotype–phenotype threshold t^{QT} for edge detection in G1 seems more sensitive and important than the linkage analysis threshold $d_{\min}$ required in preprocessing. Regarding sensitivity of the performance evaluation, AUROC is much less sensitive to the parameters t^{QT} and $d_{\min}$ than AUPR.

Larger marker distance seems beneficial for reconstruction because genotype correlations are then minimized. This can be seen, for instance, when comparing configuration 2 (marker distance N(5, 1)) with 6 (marker distance N(1, 0.1)) in Fig. 3.2. Partially, weak performance due to small marker distance can be compensated by biological variability (configuration 2 vs. 8). However, in the case of small samples and larger marker distance, larger biological variability decreases performance. This is most likely due to a poor signal-to-noise ratio and can be understood as follows. Interactions between genes are derived from target expression variations induced

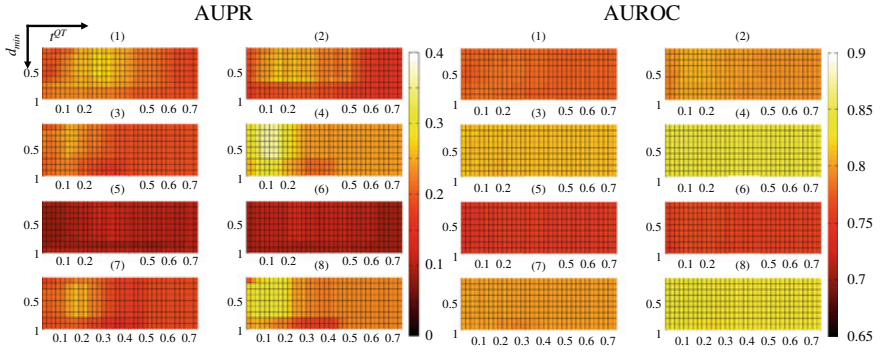


Fig. 3.2 Performance of AUPR (*left*) and AUROC (*right*) of networks 100.1.1–100.1.8 depending on the chosen threshold parameters

by regulator genotype variations. This approach requires sufficient (i) variation of the regulator and (ii) sensitivity of targets with respect to expression variations of the regulator. Variation of the regulator can only be induced by either upstream genes, i.e., the regulator itself is regulated by other genes, and/or by biological variability inducing expression variation in each gene along the sample population. The latter is important for identifying regulator–target interactions of regulators, which have no upstream genes. In this case, the only source of topological informative expression variation is biological variability, which however can only be distinguished from uninformative noise for larger sample sizes.

Figure 3.3 summarizes the AUROC and AUPR performances for all network configurations and sizes averaged over the three network replicates. These results confirm many of the observations made for networks 100.1.x. Again, for our reconstruction algorithm, the worst scenario in terms of AUPR values is the one with small sample size, small marker distance, and small biological variance. We also see that the AUROC is more or less insensitive with respect to sample size and configuration of marker distance/biological variance, but sensitive to the total number of nodes. Specifically, the AUROC is constantly decreased in networks with only 100 nodes compared to 1,000 and 5,000 nodes. This is most likely due to the fact that there are less false negative edges in small compared to large networks (if they have the same connectivity, which is the case for the given dataset) leading to a decreased AUROC. Best network configuration for reconstruction in terms of AUPR values is given by larger samples and large marker distance from which only the first one can be influenced by experimental design. Increased biological variance has noticeable effects on the reconstruction quality for small marker distance. Here, higher biological variance is favorable. The reconstruction quality with respect to network size decreases clearly in one particular case: networks with 5,000 nodes perform poorly in the AUPR values for small sample size (300). Therefore, precision is small in this setting because of too few samples. For 900 samples, precision is raised, resulting into similar AUPR values compared to reconstructions of 100/1,000 node networks.

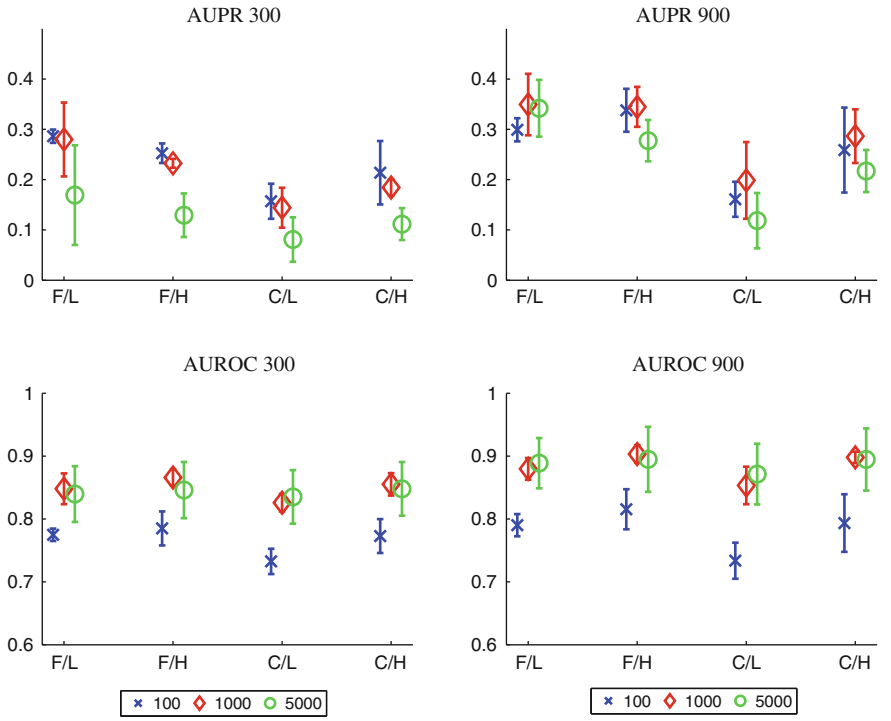
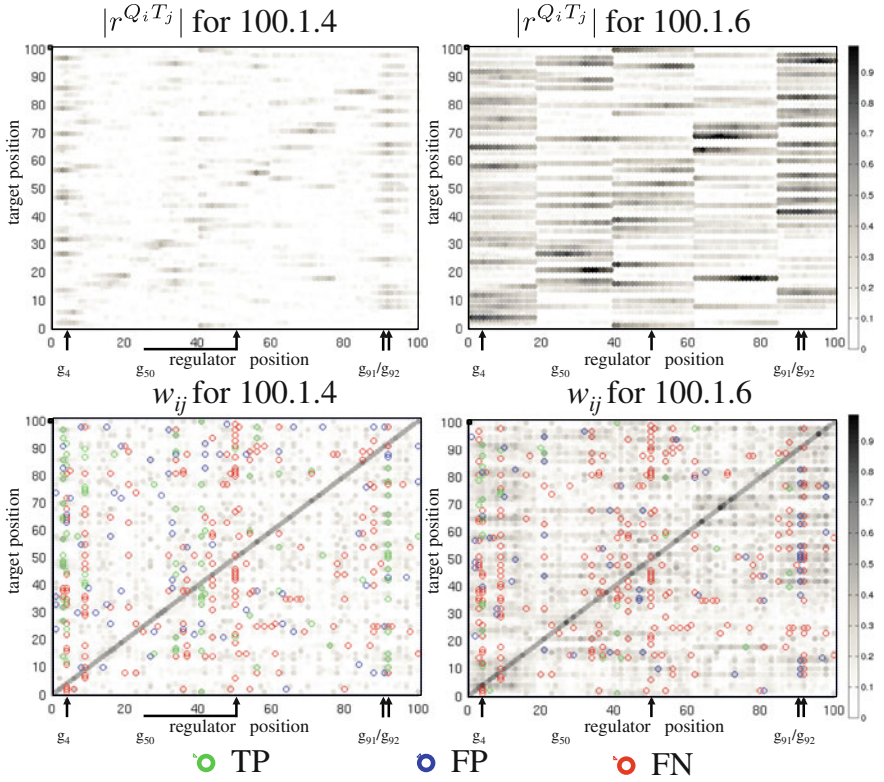


Fig. 3.3 AUPR and AUROC performance averaged over network replicates for different network sizes (100/1,000/5,000 nodes) and samples (300 (*left panel*) or 900 (*right panel*)) grouped according to marker distance (Far/Close) / biological variance (Low/High) configurations

Averaged over all configurations, networks with 1,000 nodes are best reconstructed with respect to AUPR and AUROC values for the eight different configurations.

3.3.2 Prominent Sources of Reconstruction Errors

In the following, we restrict the analysis to (i) a well-identifiable configuration (100.1.4) and (ii) a poorly identifiable configuration (100.1.6). We further restrict our analysis to 900 samples, since the influence of the sample size should be clear from the discussions above. In Fig. 3.4 we show the genotype–phenotype correlation matrix and weight matrix as a density plot. Thereby we have indicated TP (green circles), FP (blue circles), and FN (red circles) in the weight matrix (note that the green and blue circles together describe the reconstructed network G5). In the genotype–phenotype matrix plots we see horizontal gray lines (especially in 100.1.6), which correspond to eQTLs, from which regulators have to be selected, in order to reconstruct the GRN. We see that configuration 100.1.4 tends to have more



Configuration 100.1.4	Configuration 100.1.6
g_4 : 0 regulators, 26 targets	
μ_E=1.08, σ^2_E=0.19, μ_w=0.11, σ^2_w=0.009, TP=8, FP=7, FN=18	μ_E=0.82, σ^2_E=0.06, μ_w=0.21, σ^2_w=0.022, TP=12, FP=4, FN=14
g_{50} : 1 regulator, 25 targets	
μ_E=0.47, σ^2_E=0.06, μ_w=0.08, σ^2_w=0.007, TP=2, FP=0, FN=23	μ_E=0.48, σ^2_E=0.02, μ_w=0.18, σ^2_w=0.016, TP=0, FP=70 FN=25
g_{91} : 5 regulators, 0 targets	
μ_E=0.35, σ^2_E=0.1, μ_w=0.10, σ^2_w=0.014, TP=0, FP=3, FN=0	μ_E=0.4, σ^2_E=0.03, μ_w=0.19, σ^2_w=0.03, TP=0, FP=14, FN=0
g_{92} : 1 regulator, 21 targets	
μ_E=1.57, σ^2_E=0.43, μ_w=0.11, σ^2_w=0.017, TP=16, FP=5, FN=5	μ_E=1.35, σ^2_E=0.098, μ_w=0.18, σ^2_w=0.02, TP=0, FP=0, FN=21

Fig. 3.4 The *upper panel* shows the genotype–phenotype correlation matrix and the *middle panel* the edge weights (for calculation see Fig. 3.1) of all potential interactions for configurations 100.1.4/100.1.6. *Horizontal gray lines* in the genotype–phenotype correlation matrix correspond to eQTLs, from which regulator genes have to be selected. In the weight matrix, *green* (TP) and *blue circles* (FP) indicate the edges included in the final reconstructed graph G5, whereas *red circles* indicate missed interactions (FN). Some genes (g_4, g_{50}, g_{91}, and g_{92}) were selected for detailed analysis of the TP/FP/FN edges having these genes as regulators (see also Fig. 3.5). Mean expression and its variance of the regulators are given by μ_E and σ^2_E, respectively. Mean weights and weight variances over all target edges of a regulator are indicated by μ_w and σ^2_w

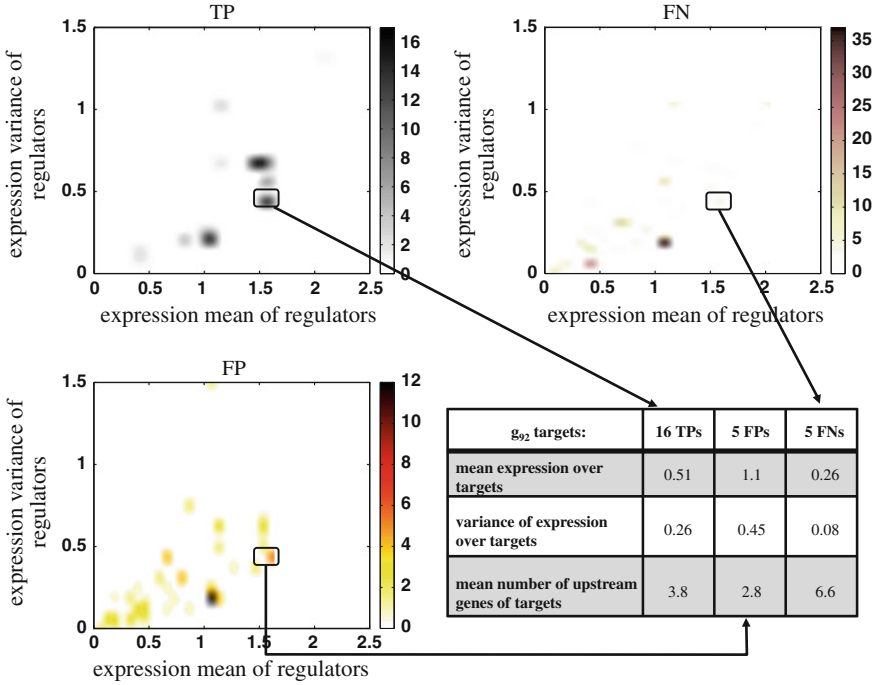
confined eQTLs due to larger marker distances, i.e., smaller genotype correlation between adjacent markers. This of course improves reconstruction quality as can be seen, e.g., in Table 3.1 (AUPR of 0.36 in 100.1.4 vs. 0.12 in 100.1.6).

From the weight matrix plots we also see that 100.1.6 contains more gray spots than 100.1.4. This results from much more correlations in the data of 100.1.6. Since many of these correlations are due to marker correlations, they do not reflect true interactions, thus hampering network inference. The diagonal gray line indicates self-regulation, which were not considered for reconstruction (and were not taken into account by the performance evaluation script). A vertical line of red or green circles indicates a true regulator with many targets. An example is regulator g_{92}, from which many targets are correctly identified in the case of 100.1.4. In the case of 100.1.6, the algorithm selects g_{91} as the regulator and therefore induces many FPs (vertical line of blue circles at regulator position 91) and many FNs (vertical line of red circles at regulator position 92). The reason for this is that eQTLs in 100.1.6 are much larger due to smaller marker distances, corresponding to a strong correlation of genes g_{91}/g_{92} via their genotypes (see genotype–phenotype matrix plot in Fig. 3.4). For configurations 100.1.4/100.1.6, gene g_{92} has 1 true upstream gene, 21 true targets, and mean expressions $\mu_E = 1.57/\mu_E = 1.35$ with ${\sigma^2}_E = 0.43/\sigma_E^2 = 0.098$. In contrast, gene g_{91} has 5 true upstream nodes, 0 true targets, and mean expressions $\mu_E = 0.35/\mu_E = 0.4$ with $\sigma_E^2 = 0.1/\sigma_E^2 = 0.03$ for configurations 100.1.4/100.1.6. Therefore, when deriving the weights for 100.1.6, gene g_{91} has larger weights with little variance than gene g_{92}, thus being wrongly selected during eQTL analysis.

Notably, even when a gene has no upstream gene (regulator), we may still recover target interactions. For example, gene g_4 has no regulator but we do recover 8 / 12 interactions out of 26 for configuration 100.1.4/100.1.6, simply due to the fact, that the expression of gene g_4 is varying due to higher biological variance resulting into expression variations of the targets (see mean edge weights of G4 targets in the table of Fig. 3.4).

Another example for typical challenges of correctly reconstructing interactions from the provided dataset is gene g_{50}. This gene has mean expressions $\mu_E = 0.47/\mu_E = 0.48$ with $\sigma_E^2 = 0.06/\sigma_E^2 = 0.02$ for configurations 100.1.4/100.1.6, with 1 true upstream gene. As the variation in the expressions of gene g_{50} is small, we cannot get any information on its targets superior to variation by noise. Further, even in cases where a regulator is varying strongly it does not necessarily induce variation in the target (see FN histogram and the table in Fig. 3.5). This can happen in cases where a gene has several regulators or if the kinetics of the target activation is in an insensitive range with respect to changes in the regulator (e.g., due to a very low or very large K_m parameter in a Hill function describing the dependency of the target on its regulator). Both effects result into small sensitivity with respect to regulators, thus hampering again the identification of interactions.

In Fig. 3.5 we show three histograms of mean and variance of the regulators' expressions, classified according to whether the (non-)identified target interactions of the regulator are TPs/FPs/FNs. We use network configuration 100.1.4 with optimal threshold parameters as it belongs to the networks with highest reconstruction quality. As expected, regions in the mean–variance expression plane in Fig. 3.5 where we



g_{92} targets:	16 TPs	5 FPs	5 FNs
mean expression over targets	0.51	1.1	0.26
variance of expression over targets	0.26	0.45	0.08
mean number of upstream genes of targets	3.8	2.8	6.6

Fig. 3.5 Histograms of expression mean versus expression variance of (non)identified regulators for network configuration 100.1.4 and classified whether the corresponding target regulation is a TP/FP/FN

find TPs also overlap with FP and FN regions. Only for mean and variance levels above 1.2 and 0.4, respectively, FNs and partially FPs are reduced. The drop in FNs is due to the fact that interactions are not missed in the high-level region of the mean–variance plane. Almost independent on the expression mean and variance of a regulator, regulators are sometimes wrongly selected from the eQTLs. This explains why FPs are only slightly reduced in the high-level region.

Interactions of regulators with expression values roughly below 0.5 and variance levels below 0.1 are always mis-classified as either FP or FN. Looking at the mean and variance of the expression levels of the target genes that belong to TP/FP/FN of regulator g_{92} (see table in Fig. 3.4), we see that sufficient variation at a sufficient expression level of the regulator does not guarantee correct identification of (no) interactions. The expression level of the target and its variance also determine classification results. The more inputs a target has, the more likely it is to get an FN since its sensitivity to variation of a specific input node is decreased (see mean expression variance over the FN target genes). False positives are also generated, when the FP targets vary too strongly. In the example of Fig. 3.5, this is probably due to strong biological variance and experimental noise, inducing variations in the FP targets; all five FP targets have a relatively low mean input number of 2.8.

3.4 Summary and Conclusions

We have analyzed the reconstruction results obtained with our recently developed framework for reconstructing gene regulatory networks based on simple correlation measures. Several different network topologies and data qualities have been used to illustrate limitations and challenges for network inference. We demonstrated that the reconstruction quality is influenced by (i) experimental design in terms of sample size and (ii) biological factors (marker distance, biological variability, and target sensitivity with respect to its regulators). Regarding the experimental design, our framework is relatively tolerant to small sample sizes, when comparing the reconstruction results from 300 and 900 sample data. However, best results are obtained with large number of samples and larger marker distances combined with significant (but not too large) biological variances. Biological factors that are beneficial for reconstruction are: larger biological variance in case of genetically close markers, input sensitivity, i.e., every gene does vary when its regulators vary in expression or genotype, respectively.

Finally, we note that meaningful reconstruction results can only be achieved when marker distances are sufficiently large. Otherwise, one should restrict the reconstruction to G3, i.e., eQTL mapping, to narrow down potential interaction sites. Then, for specific genes, the true interactions may be obtained by further focused experimental analysis based on the initial reconstructed graph G3.

References

Bing N, Hoeschele I (2005) Genetical genomic analysis of a yeast segregant population for transcription network inference. Genetics 170:533–542

Flassig RJ, Heise S, Sundmacher K, Klamt S (2013) An effective framework for reconstructing gene regulatory networks from genetical genomics data. Bioinformatics 29(2):246–254

Jansen R, Nap N (2001) Genetical genomics: the added value from segregation. Trends Genet 17:388–391

Jansen R (2003) Studying complex biological systems using multifactorial perturbation. Nat Rev Genet 4:145–151

Keurentjes JJB, Fu J, Terpstra IR et al (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci USA 104:1708–1713

Klamt S, Flassig RJ, Sundmacher K (2010) TRANSWESD: inferring cellular networks with transitive reduction. Bioinformatics 26:2160–2168

Li H, Lu L, Manly KF et al (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14:1119–1125

Liu B, de la Fuente A, Hoeschele I (2008) Gene network inference via structural equation modeling in genetical genomics experiments. Genetics 178:1763–1776

Liu B, Hoeschele I, de la Fuente A (2010) Inferring gene regulatory networks from genetical genomics data. In: Das S, Caragea D, Hsu WH, Welch SM (eds) Computational methodologies in gene regulatory networks. IGI Global, Hershey, pp 79–107

Michaelson JJ, Loguercio S, Beyer A (2009) Detection and interpretation of expression quantitative trait loci (eQTL). Methods 48:265–276

Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7:862–872
Rockman MV (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. Nature 456:738–744
Zhu J, Wiener MC, Zhang C et al (2007) Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. PLoS Comput Biol 3:e69