Supporting material (Appendix) for the thesis:

The analysis of high-throughput biological datasets utilising distributed computing

Submitted by

Jamie J. Alnasir

Supervisor: Hugh P. Shanahan

Advisor: Gregory Z. Gutin

2017

Appendix 1

Testing the Protein Databank (Chapter 5)

1.1 PDB file accessions used for benchmarking

Data used from the Protein Databank for benchmarking PDB-Hadoop (chapter 5 section 5.6) is listed below, by PDB entry (accession):

```
pdb1a30, pdb1a31, pdb1a32, pdb1a33, pdb1a34, pdb1a35, pdb1a36, pdb1a37, pdb1a38,
pdb1a39, pdb1a3a, pdb1a3b, pdb1a3c, pdb1a3d, pdb1a3e, pdb1a3f, pdb1a3g, pdb1a3h,
pdb1a3i, pdb1a3j, pdb1a3k, pdb1a3l, pdb1a3m, pdb1a3n, pdb1a3o, pdb1a3p, pdb1a3q,
pdb1a3r, pdb1a3s, pdb1a3t, pdb1a3u, pdb1a3v, pdb1a3w, pdb1a3x, pdb1a3y, pdb1a3z,
pdb1dk0, pdb1dk1, pdb1dk2, pdb1dk3, pdb1dk4, pdb1dk5, pdb1dk6, pdb1dk7, pdb1dk8,
pdb1dk9, pdb1dka, pdb1dkc, pdb1dkd, pdb1dke, pdb1dkf, pdb1dkg, pdb1dkh, pdb1dki,
pdb1dkj, pdb1dkk, pdb1dkl, pdb1dkm, pdb1dkn, pdb1dko, pdb1dkp, pdb1dkq, pdb1dkr,
pdb1dks, pdb1dkt, pdb1dku, pdb1dkw, pdb1dkx, pdb1dky, pdb1dkz, pdb1e50, pdb1e51,
pdb1e52, pdb1e54, pdb1e55, pdb1e56, pdb1e57, pdb1e58, pdb1e59, pdb1e5a, pdb1e5b,
pdb1e5c, pdb1e5d, pdb1e5e, pdb1e5f, pdb1e5g, pdb1e5h, pdb1e5i, pdb1e5j, pdb1e5k,
pdb1e5l, pdb1e5m, pdb1e5n, pdb1e5o, pdb1e5p, pdb1e5q, pdb1e5r, pdb1e5s, pdb1e5t,
pdb1e5u, pdb1e5v, pdb1e5w, pdb1e5x, pdb1e5y, pdb1e5z, pdb1ef0, pdb1ef1, pdb1ef2,
pdb1ef3, pdb1ef4, pdb1ef5, pdb1ef7, pdb1ef8, pdb1ef9, pdb1efa, pdb1efc, pdb1efd,
pdb1efe, pdb1efg, pdb1efh, pdb1efi, pdb1efk, pdb1efl, pdb1efm, pdb1efn, pdb1efo,
pdb1efp, pdb1efq, pdb1efr, pdb1efs, pdb1eft, pdb1efu, pdb1efv, pdb1efw, pdb1efx,
pdb1efy, pdb1efz, pdb1fj0, pdb1fj1, pdb1fj2, pdb1fj3, pdb1fj4, pdb1fj5, pdb1fj6,
pdb1fj7, pdb1fj8, pdb1fj9, pdb1fja, pdb1fjb, pdb1fjc, pdb1fjd, pdb1fje, pdb1fjg,
pdb1fjh, pdb1fjj, pdb1fjk, pdb1fjl, pdb1fjm, pdb1fjn, pdb1fjo, pdb1fjp, pdb1fjq,
pdb1fjr, pdb1fjs, pdb1fjt, pdb1fju, pdb1fjv, pdb1fjw, pdb1fjx, pdb1fp0, pdb1fp1,
pdb1fp2, pdb1fp3, pdb1fp4, pdb1fp5, pdb1fp6, pdb1fp7, pdb1fp8, pdb1fp9, pdb1fpb,
pdb1fpc, pdb1fpd, pdb1fpe, pdb1fpf, pdb1fpg, pdb1fph, pdb1fpi, pdb1fpj, pdb1fpk,
pdb1fpl, pdb1fpm, pdb1fpn, pdb1fpo, pdb1fpp, pdb1fpq, pdb1fpr, pdb1fps, pdb1fpt,
pdb1fpu, pdb1fpv, pdb1fpw, pdb1fpx, pdb1fpy, pdb1fpz, pdb1ga0, pdb1ga1, pdb1ga2,
pdb1ga3, pdb1ga4, pdb1ga5, pdb1ga6, pdb1ga7, pdb1ga8, pdb1ga9, pdb1gab, pdb1gac,
pdb1gad, pdb1gae, pdb1gaf, pdb1gag, pdb1gah, pdb1gai, pdb1gaj, pdb1gak, pdb1gal,
pdb1gam, pdb1gan, pdb1gao, pdb1gaq, pdb1gar, pdb1gat, pdb1gau, pdb1gav, pdb1gaw,
pdb1gax, pdb1gay, pdb1gaz, pdb1uq4, pdb1uq5, pdb1uqa, pdb1uqb, pdb1uqc, pdb1uqd,
pdb1uqe, pdb1uqf, pdb1uqg, pdb1uqr, pdb1uqs, pdb1uqt, pdb1uqu, pdb1uqv, pdb1uqw,
pdb1uqx, pdb1uqy, pdb1uqz, pdb2a30, pdb2a31, pdb2a32, pdb2a33, pdb2a35, pdb2a36,
pdb2a37, pdb2a38, pdb2a39, pdb2a3a, pdb2a3b, pdb2a3c, pdb2a3d, pdb2a3e, pdb2a3f,
```

pdb2a3g, pdb2a3h, pdb2a3i, pdb2a3j, pdb2a3k, pdb2a3l, pdb2a3m, pdb2a3n, pdb2a3p,
pdb2a3q, pdb2a3r, pdb2a3s, pdb2a3t, pdb2a3u, pdb2a3v, pdb2a3w, pdb2a3x, pdb2a3y,
pdb2a3z, pdb2dk1, pdb2dk2, pdb2dk3, pdb2dk4, pdb2dk5, pdb2dk6, pdb2dk7, pdb2dk8,
pdb2dk9, pdb2dka, pdb2dkb, pdb2dkc, pdb2dkd, pdb2dke, pdb2dkf, pdb2dkg, pdb2dkh,
pdb2dki, pdb2dkj, pdb2dkk, pdb2dkl, pdb2dkm, pdb2dkn, pdb2dko, pdb2dkp, pdb2dkq,
pdb2dkr, pdb2dks, pdb2dkt, pdb2dku, pdb2dkv, pdb2dkw, pdb2dkx, pdb2dky, pdb2dkz,
pdb2e50, pdb2e51, pdb2e52, pdb2e53, pdb2e54, pdb2e55, pdb2e56, pdb2e59, pdb2e5a,
pdb2e5b, pdb2e5c, pdb2e5d, pdb2e5e, pdb2e5f, pdb2e5g, pdb2e5h, pdb2e5i, pdb2e5j,
pdb2e5k, pdb2e5l, pdb2e5m, pdb2e5n, pdb2e5o, pdb2e5p, pdb2e5q, pdb2e5r, pdb2e5s,
pdb2e5t, pdb2e5u, pdb2e5v, pdb2e5w, pdb2e5x, pdb2e5y, pdb2e5z, pdb2ef0, pdb2ef1,
pdb2ef2, pdb2ef4, pdb2ef5, pdb2ef6, pdb2ef7, pdb2ef8, pdb2ef9, pdb2efa, pdb2efb,
pdb2efc, pdb2efd, pdb2efe, pdb2eff, pdb2efg, pdb2efh, pdb2efi, pdb2efj, pdb2efk,
pdb2efl, pdb2efn, pdb2efo, pdb2efp, pdb2efq, pdb2efr, pdb2efs, pdb2eft, pdb2efu,
pdb2efv, pdb2efw, pdb2efx, pdb2efy, pdb2efz, pdb2fj0, pdb2fj1, pdb2fj2, pdb2fj3,
pdb2fj4, pdb2fj5, pdb2fj6, pdb2fj7, pdb2fj8, pdb2fj9, pdb2fja, pdb2fjb, pdb2fjc,
pdb2fjd, pdb2fje, pdb2fjf, pdb2fjg, pdb2fjh, pdb2fji, pdb2fjk, pdb2fjl, pdb2fjm,
pdb2fjn, pdb2fjp, pdb2fjr, pdb2fjs, pdb2fjt, pdb2fju, pdb2fjv, pdb2fjw, pdb2fjx,
pdb2fjy, pdb2fjz, pdb2fp0, pdb2fp1, pdb2fp2, pdb2fp3, pdb2fp4, pdb2fp7, pdb2fp8,
pdb2fp9, pdb2fpb, pdb2fpc, pdb2fpd, pdb2fpe, pdb2fpf, pdb2fpg, pdb2fph, pdb2fpi,
pdb2fpk, pdb2fpl, pdb2fpm, pdb2fpn, pdb2fpo, pdb2fpp, pdb2fpq, pdb2fpr, pdb2fps,
pdb2fpt, pdb2fpu, pdb2fpv, pdb2fpw, pdb2fpx, pdb2fpy, pdb2fpz, pdb2ga0, pdb2ga1,
pdb2ga2, pdb2ga3, pdb2ga4, pdb2ga5, pdb2ga6, pdb2ga7, pdb2ga8, pdb2ga9, pdb2gaa,
pdb2gab, pdb2gac, pdb2gae, pdb2gaf, pdb2gag, pdb2gah, pdb2gai, pdb2gaj, pdb2gak,
pdb2gal, pdb2gam, pdb2gan, pdb2gao, pdb2gaq, pdb2gar, pdb2gas, pdb2gat, pdb2gau,
pdb2gaw, pdb2gax, pdb2gaz, pdb3a30, pdb3a31, pdb3a32, pdb3a33, pdb3a34, pdb3a35,
pdb3a36, pdb3a37, pdb3a38, pdb3a39, pdb3a3a, pdb3a3b, pdb3a3c, pdb3a3d, pdb3a3e,
pdb3a3f, pdb3a3g, pdb3a3h, pdb3a3i, pdb3a3j, pdb3a3k, pdb3a3n, pdb3a3o, pdb3a3p,
pdb3a3q, pdb3a3r, pdb3a3t, pdb3a3u, pdb3a3v, pdb3a3w, pdb3a3x, pdb3a3y, pdb3a3z,
pdb3dk0, pdb3dk1, pdb3dk2, pdb3dk3, pdb3dk4, pdb3dk5, pdb3dk6, pdb3dk7, pdb3dk8,
pdb3dk9, pdb3dka, pdb3dkb, pdb3dkc, pdb3dkd, pdb3dke, pdb3dkf, pdb3dkg, pdb3dkh,
pdb3dki, pdb3dkj, pdb3dkk, pdb3dkl, pdb3dkm, pdb3dkn, pdb3dko, pdb3dkp, pdb3dkq,
pdb3dkr, pdb3dks, pdb3dkt, pdb3dku, pdb3dkv, pdb3dkw, pdb3dkx, pdb3dky, pdb3dkz,
pdb3e50, pdb3e51, pdb3e53, pdb3e54, pdb3e55, pdb3e56, pdb3e57, pdb3e58, pdb3e59,
pdb3e5a, pdb3e5b, pdb3e5c, pdb3e5d, pdb3e5e, pdb3e5f, pdb3e5h, pdb3e5i, pdb3e5j,
pdb3e5k, pdb3e5l, pdb3e5m, pdb3e5n, pdb3e5o, pdb3e5p, pdb3e5q, pdb3e5r, pdb3e5s,
pdb3e5t, pdb3e5u, pdb3e5v, pdb3e5w, pdb3e5x, pdb3e5y, pdb3e5z, pdb3ef0, pdb3ef1,
pdb3ef2, pdb3ef3, pdb3ef4, pdb3ef5, pdb3ef6, pdb3ef7, pdb3ef8, pdb3ef9, pdb3efa,
pdb3efb, pdb3efc, pdb3efd, pdb3efe, pdb3eff, pdb3efg, pdb3efh, pdb3efi, pdb3efj,
pdb3efk, pdb3efl, pdb3efm, pdb3efo, pdb3efp, pdb3efq, pdb3efr, pdb3efs, pdb3eft,
pdb3efu, pdb3efv, pdb3efw, pdb3efx, pdb3efy, pdb3efz, pdb3fj1, pdb3fj2, pdb3fj4,
pdb3fj5, pdb3fj6, pdb3fj7, pdb3fj8, pdb3fj9, pdb3fja, pdb3fjb, pdb3fjc, pdb3fjd,
pdb3fje, pdb3fjf, pdb3fjg, pdb3fjh, pdb3fji, pdb3fjj, pdb3fjk, pdb3fjl, pdb3fjm,
pdb3fjn, pdb3fjo, pdb3fjp, pdb3fjq, pdb3fjs, pdb3fjt, pdb3fju, pdb3fjv, pdb3fjw,
pdb3fjx, pdb3fjy, pdb3fjz, pdb3fp0, pdb3fp2, pdb3fp3, pdb3fp4, pdb3fp5, pdb3fp6,
pdb3fp7, pdb3fp8, pdb3fp9, pdb3fpa, pdb3fpb, pdb3fpc, pdb3fpd, pdb3fpe, pdb3fpf,
pdb3fpg, pdb3fph, pdb3fpi, pdb3fpj, pdb3fpk, pdb3fpl, pdb3fpm, pdb3fpn, pdb3fpo,
pdb3fpp, pdb3fpq, pdb3fpr, pdb3fps, pdb3fpt, pdb3fpu, pdb3fpv, pdb3fpw, pdb3fpx,
pdb3fpy, pdb3fpz, pdb3ga0, pdb3ga1, pdb3ga2, pdb3ga3, pdb3ga4, pdb3ga5, pdb3ga6,
pdb3ga7, pdb3ga8, pdb3ga9, pdb3gaa, pdb3gab, pdb3gac, pdb3gad, pdb3gae, pdb3gaf,
pdb3gag, pdb3gah, pdb3gai, pdb3gaj, pdb3gak, pdb3gal, pdb3gam, pdb3gan, pdb3gao,
pdb3gaq, pdb3gar, pdb3gas, pdb3gat, pdb3gau, pdb3gav, pdb3gaw, pdb3gax, pdb3gay,
pdb3gaz, pdb3uq0, pdb3uq2, pdb3uq3, pdb3uq4, pdb3uq5, pdb3uq6, pdb3uq7, pdb3uq8,
pdb3uq9, pdb3uqa, pdb3uqb, pdb3uqc, pdb3uqd, pdb3uqe, pdb3uqf, pdb3uqg, pdb3uqh,
pdb3uqi, pdb3uqn, pdb3uqo, pdb3uqp, pdb3uqr, pdb3uqs, pdb3uqu, pdb3uqv, pdb3uqw,
pdb3uqx, pdb3uqy, pdb3uqz, pdb3zz1, pdb4a30, pdb4a31, pdb4a32, pdb4a33, pdb4a34,

pdb4a35, pdb4a36, pdb4a37, pdb4a38, pdb4a39, pdb4a3b, pdb4a3c, pdb4a3d, pdb4a3e,
pdb4a3f, pdb4a3g, pdb4a3h, pdb4a3i, pdb4a3j, pdb4a3k, pdb4a3l, pdb4a3m, pdb4a3n,
pdb4a3o, pdb4a3p, pdb4a3q, pdb4a3r, pdb4a3s, pdb4a3t, pdb4a3u, pdb4a3v, pdb4a3w,
pdb4a3x, pdb4a3y, pdb4a3z, pdb4dk0, pdb4dk1, pdb4dk2, pdb4dk3, pdb4dk4, pdb4dk5,
pdb4dk6, pdb4dk7, pdb4dk8, pdb4dk9, pdb4dka, pdb4dkb, pdb4dkc, pdb4dkd, pdb4dke,
pdb4dkf, pdb4dki, pdb4dkj, pdb4dkk, pdb4dkl, pdb4dkm, pdb4dkn, pdb4dko, pdb4dkp,
pdb4dkq, pdb4dkr, pdb4dks, pdb4dkt, pdb4dku, pdb4dkv, pdb4dkw, pdb4dkx, pdb4dky,
pdb4dkz, pdb4e50, pdb4e51, pdb4e52, pdb4e53, pdb4e54, pdb4e55, pdb4e56, pdb4e57,
pdb4e58, pdb4e59, pdb4e5a, pdb4e5b, pdb4e5c, pdb4e5d, pdb4e5e, pdb4e5f, pdb4e5g,
pdb4e5h, pdb4e5i, pdb4e5j, pdb4e5k, pdb4e5l, pdb4e5m, pdb4e5n, pdb4e5o, pdb4e5p,
pdb4e5q, pdb4e5r, pdb4e5s, pdb4e5t, pdb4e5u, pdb4e5v, pdb4e5w, pdb4e5x, pdb4e5y,
pdb4e5z, pdb4ef0, pdb4ef1, pdb4ef2, pdb4ef3, pdb4ef4, pdb4ef5, pdb4ef6, pdb4ef8,
pdb4ef9, pdb4efa, pdb4efb, pdb4efc, pdb4efd, pdb4efe, pdb4eff, pdb4efg, pdb4efh,
pdb4efi, pdb4efj, pdb4efk, pdb4efl, pdb4efm, pdb4efn, pdb4efo, pdb4efp, pdb4efq,
pdb4efr, pdb4efs, pdb4eft, pdb4efu, pdb4efv, pdb4efx, pdb4efz, pdb4fj0, pdb4fj1,
pdb4fj2, pdb4fj3, pdb4fj4, pdb4fj5, pdb4fj6, pdb4fj7, pdb4fj8, pdb4fj9, pdb4fjc,
pdb4fjg, pdb4fjh, pdb4fji, pdb4fjj, pdb4fjk, pdb4fjl, pdb4fjm, pdb4fjn, pdb4fjo,
pdb4fjp, pdb4fjq, pdb4fjr, pdb4fjs, pdb4fju, pdb4fjv, pdb4fjw, pdb4fjx, pdb4fjy,
pdb4fjz, pdb4fp1, pdb4fp2, pdb4fp3, pdb4fp4, pdb4fp5, pdb4fp6, pdb4fp7, pdb4fp8,
pdb4fp9, pdb4fpa, pdb4fpb, pdb4fpc, pdb4fpd, pdb4fpe, pdb4fpf, pdb4fpg, pdb4fph,
pdb4fpi, pdb4fpj, pdb4fpk, pdb4fpl, pdb4fpo, pdb4fpp, pdb4fpr, pdb4fps, pdb4fpt,
pdb4fpv, pdb4fpw, pdb4fpx, pdb4fpy, pdb4fpz, pdb4ga0, pdb4ga1, pdb4ga2, pdb4ga3,
pdb4ga4, pdb4ga5, pdb4ga6, pdb4ga7, pdb4ga8, pdb4ga9, pdb4gaa, pdb4gab, pdb4gac,
pdb4gad, pdb4gae, pdb4gaf, pdb4gag, pdb4gah, pdb4gai, pdb4gaj, pdb4gak, pdb4gal,
pdb4gam, pdb4gao, pdb4gap, pdb4gaq, pdb4gar, pdb4gas, pdb4gat, pdb4gau, pdb4gav,
pdb4gaw, pdb4gax, pdb4gay, pdb4gaz, pdb5a3h, pdb5gal, pdb5gat, pdb6a3h, pdb6gat,
pdb7a3h

1.2 Sample PDB-Hadoop job output

1.2.1 Dihedrals job output

A sample of the output from PDB-Hadoop during the computation of torsional (dihedral) angles on the PDB macro-molecular structure entries listed in section 1.1 above (discussed in chapter 5, section 5.5.1). NB: Only the first 20 records in the PDB is shown for two entries.

```
Extracted/writing file /tmp/pdb3gax.ent
pdb3gax.ent-0000000001 Phi,Psi,Omega,Chain,Residue
pdb3gax.ent-0000000002 1, 0.00, 164.35, 179.04, "A", "GLY12"
pdb3gax.ent-0000000003 2, -71.04, 145.47, -179.55, "A", "PRO13"
pdb3gax.ent-0000000004 3, -107.27, 154.17, 175.62, "A", "MET14"
pdb3gax.ent-0000000005 4, -58.21, 144.81, 175.34, "A", "ASP15"
pdb3gax.ent-0000000006 5, -141.64, 175.10, -178.81, "A", "ALA16"
pdb3gax.ent-0000000007 6, -143.17, 162.72, 177.07, "A", "SER17"
pdb3gax.ent-0000000008 7, -68.77, -13.42, 167.40, "A", "VAL18"
pdb3gax.ent-0000000009 8, -63.05, -22.24, -179.18, "A", "GLU19"
pdb3gax.ent-0000000010 9, -76.13, 128.76, -179.80, "A", "GLU20"
pdb3gax.ent-0000000011 10, -53.96, -43.56, 179.65, "A", "GLU21"
pdb3gax.ent-0000000012 11, -62.53, -44.40, 172.47, "A", "GLY22"
pdb3gax.ent-0000000013 12, -55.75, -47.94, 177.52, "A", "VAL23"
pdb3gax.ent-0000000014 13, -59.85, -41.82, 179.37, "A", "ARG24"
pdb3gax.ent-0000000015 14, -64.57, -44.26, 177.91, "A", "ARG25"
pdb3gax.ent-0000000016 15, -64.42, -39.67, 176.43, "A", "ALA26"
```

```
pdb3gax.ent-0000000017 16, -59.85, -49.39, -179.06, "A", "LEU27"
pdb3gax.ent-0000000018 17, -66.68, -41.04, 173.57, "A", "ASP28"
pdb3gax.ent-0000000019 18, -55.45, -52.34, -176.12, "A", "PHE29"
pdb3gax.ent-0000000020 19, -62.26, -45.43, 178.14, "A", "ALA30"
pdb3gax.ent-0000000021 20, -65.31, -41.39, 172.49, "A", "VAL31"
......
......
Extracted/writing file /tmp/pdb3zzs.ent
pdb3ga3.ent-0000000001 Phi,Psi,Omega,Chain,Residue
pdb3ga3.ent-0000000002 1, 0.00, -25.54, -177.36, "A", "ALA893"
pdb3ga3.ent-0000000003 2, -63.68, -21.97, 178.11, "A", "LYS894"
pdb3ga3.ent-0000000004 3, -103.02, 7.32, 178.37, "A", "HIS895"
pdb3ga3.ent-0000000005 4, -68.56, 160.12, 177.24, "A", "TYR896"
pdb3ga3.ent-0000000006 5, -92.42, 128.71, 179.37, "A", "LYS897"
pdb3ga3.ent-0000000007 6, -94.55, 50.91, -178.00, "A", "ASN898"
pdb3ga3.ent-0000000008 7, -133.32, 78.56, -179.84, "A", "ASN899"
pdb3ga3.ent-0000000009 8, -60.43, -21.70, 179.64, "A", "PRO900"
pdb3ga3.ent-0000000010 9, -63.02, -13.94, 179.82, "A", "SER901"
pdb3ga3.ent-0000000011 10, -70.68, -17.54, -176.86, "A", "LEU902"
pdb3ga3.ent-0000000012 11, -122.34, 157.51, 175.84, "A", "ILE903"
pdb3ga3.ent-0000000013 12, -132.36, 149.02, 179.17, "A", "THR904"
pdb3ga3.ent-0000000014 13, -111.26, 126.57, 178.44, "A", "PHE905"
pdb3ga3.ent-0000000015 14, -113.43, 144.65, 175.13, "A", "LEU906"
pdb3ga3.ent-0000000016 15, -58.98, 131.57, -177.90, "A", "CYS907"
pdb3ga3.ent-0000000017 16, -66.44, -22.86, -179.84, "A", "LYS908"
pdb3ga3.ent-0000000018 17, -89.35, -53.63, -176.44, "A", "ASN909"
pdb3ga3.ent-0000000019 18, -101.81, -4.60, 178.96, "A", "CYS910"
pdb3ga3.ent-0000000020 19, 69.58, 7.36, 179.63, "A", "SER911"
pdb3ga3.ent-0000000021 20, -68.24, 147.11, 177.33, "A", "VAL912"
......
......
```

1.2.2 Docking job output

Below is a sample of the output from PDB-Hadoop during the docking of a small oligo-peptide (discussed in chapter 5, section 5.5.2) against the PDB macro-molecular entries listed in section 1.1 above. The output employs the *post-processing* step of PDB-Hadoop to extract docking scores from Vina AutoDock and summarise them in order of best (lowest energy) docking score.

```
Extracted/writing file /tmp/pdb1a3q.ent
Initiating post-processing...
pdb1a3q.ent-0000000001     1         -4.6      0.000      0.000
pdb1a3q.ent-0000000002     2         -4.4     22.399     23.769
pdb1a3q.ent-0000000003     3         -4.2     22.208     22.970
pdb1a3q.ent-0000000004     4         -4.2     19.226     21.204
pdb1a3q.ent-0000000005     5         -4.1     34.622     36.402
pdb1a3q.ent-0000000006     6         -4.0     20.313     22.641
pdb1a3q.ent-0000000007     7         -4.0     34.348     35.354
pdb1a3q.ent-0000000008     8         -4.0     27.283     29.148
pdb1a3q.ent-0000000009     9         -3.9     26.239     27.588
Extracted/writing file /tmp/pdb2dkr.ent
Initiating post-processing...
pdb2dkr.ent-0000000001     1         -5.0      0.000      0.000
```

```
pdb2dkr.ent-0000000002     2         -4.7      3.496      5.931
pdb2dkr.ent-0000000003     3         -4.6      3.759      4.530
pdb2dkr.ent-0000000004     4         -4.6      5.178      6.929
pdb2dkr.ent-0000000005     5         -4.4     24.758     26.165
pdb2dkr.ent-0000000006     6         -4.3      2.661      3.299
pdb2dkr.ent-0000000007     7         -4.2     21.788     22.757
pdb2dkr.ent-0000000008     8         -4.0     17.432     19.035
pdb2dkr.ent-0000000009     9         -4.0     14.814     15.540
......
......
```

Appendix 2

Transcriptomics analysis (Chapter 6)

2.1 Wild-type *D. melanogaster* results

Motif spacing: **10bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	8	96
p(t-test)	$4.83\text{x}10^{-4}(1.54\text{x}10^{-3})$*		$9.91\text{x}10^{-1}(9.91\text{x}10^{-1})$		$6.56\text{x}10^{-1}(7.50\text{x}10^{-1})$		$3.50\text{x}10^{-1}(4.38\text{x}10^{-1})$	
p(Wilcoxon)	$8.79\text{x}10^{-2}(1.41\text{x}10^{-1})$		$1.96\text{x}10^{-1}(2.61\text{x}10^{-1})$		$4.69\text{x}10^{-1}(5.36\text{x}10^{-1})$		1.00(1.00)	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	54	96
p(t-test)	$7.66\text{x}10^{-3}(1.53\text{x}10^{-2})$*		$8.53\text{x}10^{-1}(9.10\text{x}10^{-1})$		$3.31\text{x}10^{-1}(4.38\text{x}10^{-1})$		$4.81\text{x}10^{-2}(8.55\text{x}10^{-2})$	
p(Wilcoxon)	$5.40\text{x}10^{-3}(1.73\text{x}10^{-2})$*		$9.52\text{x}10^{-1}(1.00)$		$1.83\text{x}10^{-1}(2.61\text{x}10^{-1})$		$8.58\text{x}10^{-2}(1.41\text{x}10^{-1})$	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	$1.40\text{x}10^{-1}(2.25\text{x}10^{-1})$		$3.55\text{x}10^{-1}(4.38\text{x}10^{-1})$		$6.26\text{x}10^{-3}(1.43\text{x}10^{-2})$*		$9.88\text{x}10^{-5}(3.95\text{x}10^{-4})$*	
p(Wilcoxon)	$2.30\text{x}10^{-2}(4.60\text{x}10^{-2})$*		$4.11\text{x}10^{-1}(5.06\text{x}10^{-1})$		$1.00\text{x}10^{-2}(2.68\text{x}10^{-2})$*		$8.07\text{x}10^{-4}(1.28\text{x}10^{-2})$*	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	$1.08\text{x}10^{-3}(2.89\text{x}10^{-3})$*		$4.82\text{x}10^{-6}(2.57\text{x}10^{-5})$*		$2.28\text{x}10^{-6}(1.82\text{x}10^{-5})$*		$3.45\text{x}10^{-8}(5.52\text{x}10^{-7})$*	
p(Wilcoxon)	$1.31\text{x}10^{-2}(2.99\text{x}10^{-2})$*		$2.71\text{x}10^{-3}(1.28\text{x}10^{-2})$*		$1.92\text{x}10^{-3}(1.28\text{x}10^{-2})$*		$3.20\text{x}10^{-3}(1.28\text{x}10^{-2})$*	

Table 2.1: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 10bp spacing for varying motif GC and mean exon GC content in Wild-type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

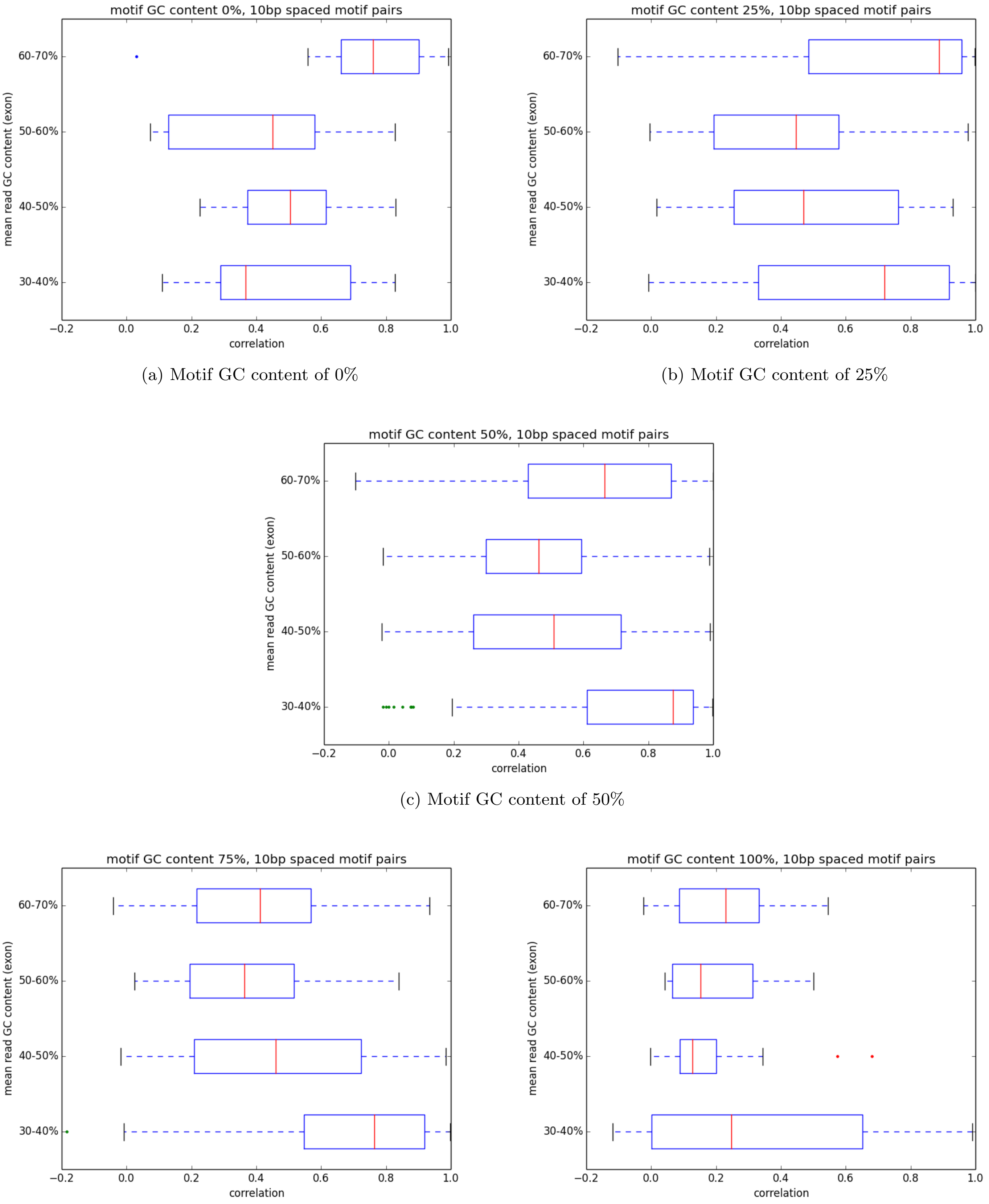
Motif spacing: **50bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	4	96
p(t-test)	2.87x10^{-2}(7.07x10^{-2})		4.94x10^{-1}(6.32x10^{-1})		9.91x10^{-1}(9.91x10^{-1})		5.58x10^{-1}(6.87x10^{-1})	
p(Wilcoxon)	5.23x10^{-3}(2.47x10^{-2})*		4.38x10^{-1}(5.84x10^{-1})		5.35x10^{-1}(6.58x10^{-1})		1.00(1.00)	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	46	96
p(t-test)	3.89x10^{-1}(5.42x10^{-1})		4.87x10^{-1}(6.32x10^{-1})		8.69x10^{-1}(9.31x10^{-1})		1.63x10^{-1}(2.75x10^{-1})	
p(Wilcoxon)	1.58x10^{-1}(2.81x10^{-1})		3.29x10^{-1}(4.78x10^{-1})		8.94x10^{-1}(9.86x10^{-1})		7.97x10^{-1}(9.11x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	61	96	64	96	64	96	64	96
p(t-test)	8.73x10^{-1}(9.31x10^{-1})		1.15x10^{-1}(2.17x10^{-1})		7.15x10^{-2}(1.51x10^{-1})		7.53x10^{-2}(1.51x10^{-1})	
p(Wilcoxon)	7.82x10^{-1}(9.11x10^{-1})		5.49x10^{-2}(1.17x10^{-1})		2.04x10^{-1}(3.11x10^{-1})		4.30x10^{-3}(2.47x10^{-2})*	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	5.69x10^{-3}(2.00x10^{-2})*		1.06x10^{-2}(2.82x10^{-2})*		1.36x10^{-3}(5.42x10^{-3})*		7.23x10^{-5}(5.78x10^{-4})*	
p(Wilcoxon)	9.73x10^{-3}(3.21x10^{-2})*		1.74x10^{-2}(4.63x10^{-2})*		8.36x10^{-3}(3.21x10^{-2})*		2.00x10^{-2}(4.92x10^{-2})*	

Table 2.2: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 50bp spacing for varying motif GC and mean exon GC content in Wild-type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

Motif spacing: **100bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	3	96
p(t-test)	1.31x10^{-2}(3.14x10^{-2})*		2.03x10^{-2}(4.64x10^{-2})*		4.54x10^{-1}(5.73x10^{-1})		1.06x10^{-4}(5.66x10^{-4})*	
p(Wilcoxon)	8.79x10^{-2}(1.62x10^{-1})		2.78x10^{-1}(3.92x10^{-1})		6.05x10^{-1}(6.92x10^{-1})		1.09x10^{-1}(1.87x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	52	96
p(t-test)	1.31x10^{-1}(2.17x10^{-1})		4.58x10^{-2}(9.16x10^{-2})		4.51x10^{-2}(9.16x10^{-2})		5.57x10^{-1}(6.38x10^{-1})	
p(Wilcoxon)	8.09x10^{-2}(1.62x10^{-1})		1.73x10^{-2}(4.63x10^{-2})*		1.06x10^{-1}(1.87x10^{-1})		4.39x10^{-1}(5.40x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	3.23x10^{-1}(4.68x10^{-1})		6.12x10^{-3}(1.77x10^{-2})*		3.11x10^{-1}(4.67x10^{-1})		1.28x10^{-3}(4.65x10^{-3})*	
p(Wilcoxon)	1.45x10^{-1}(2.40x10^{-1})		3.46x10^{-2}(7.90x10^{-2})		3.99x10^{-1}(5.33x10^{-1})		3.26x10^{-3}(2.23x10^{-2})*	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	5.91x10^{-4}(2.58x10^{-3})*		2.93x10^{-9}(7.03x10^{-8})*		1.84x10^{-6}(2.19x10^{-5})*		4.67E-11(2.24x10^{-9})*	
p(Wilcoxon)	1.51x10^{-2}(4.53x10^{-2})*		1.12x10^{-3}(1.80x10^{-2})*		1.31x10^{-2}(4.18x10^{-2})*		7.76x10^{-4}(1.80x10^{-2})*	

Table 2.3: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 100bp spacing for varying motif GC and mean exon GC content in Wild-type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

Wild type *D. melanogaster* - motif-pair correlations at 10bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

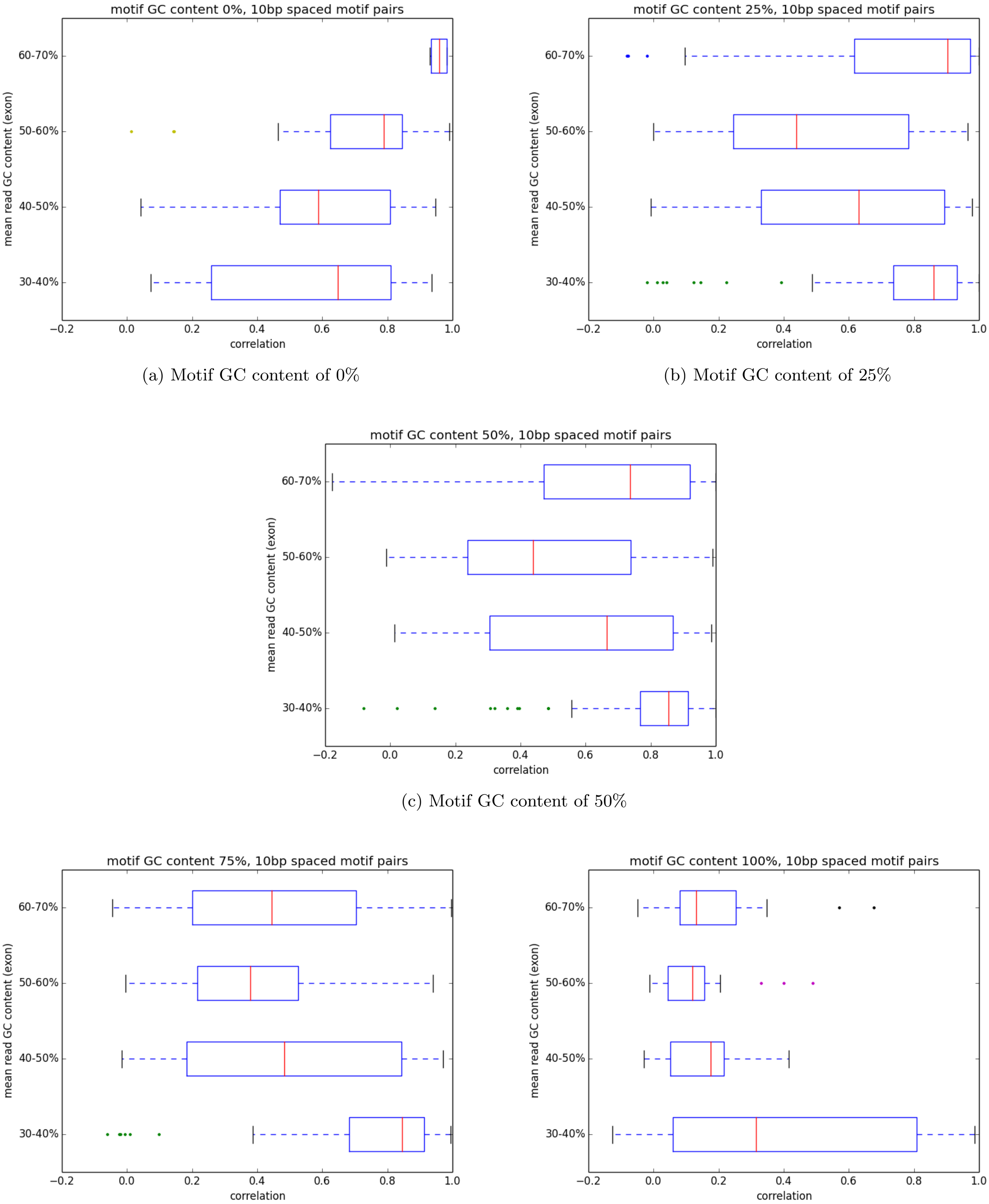
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.1: Box and whisker plots of motif-pair correlations at a distance of 10bp for Wild-type *D. melanogaster*

Wild type *D. melanogaster* - motif pair correlations at 10bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

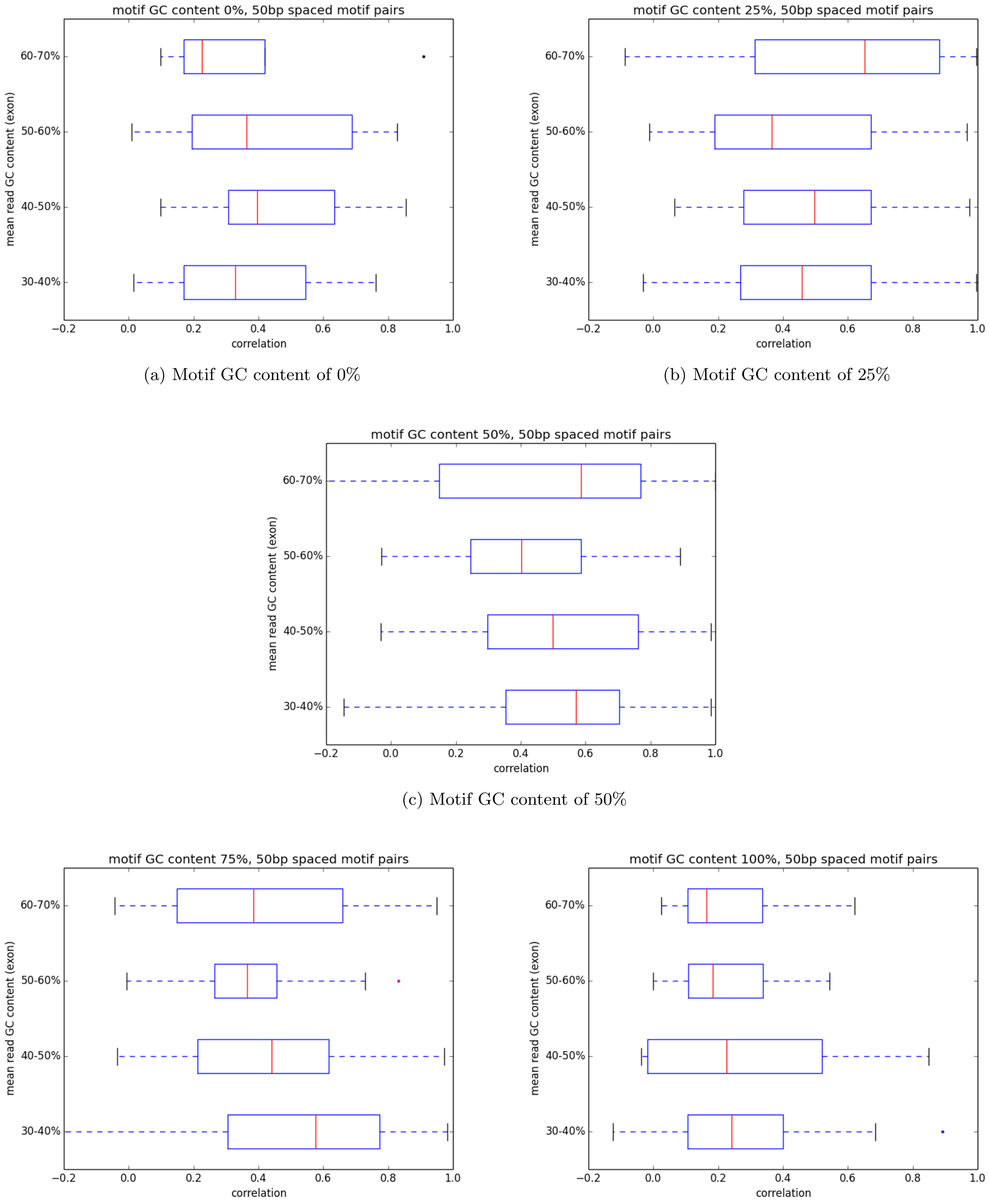
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.2: Box and whisker plots of motif pair correlations at a distance of 10bp for Wild-type *D. melanogaster* (Excluding hexamer regions)

Wild type *D. melanogaster* - motif-pair correlations at 50bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

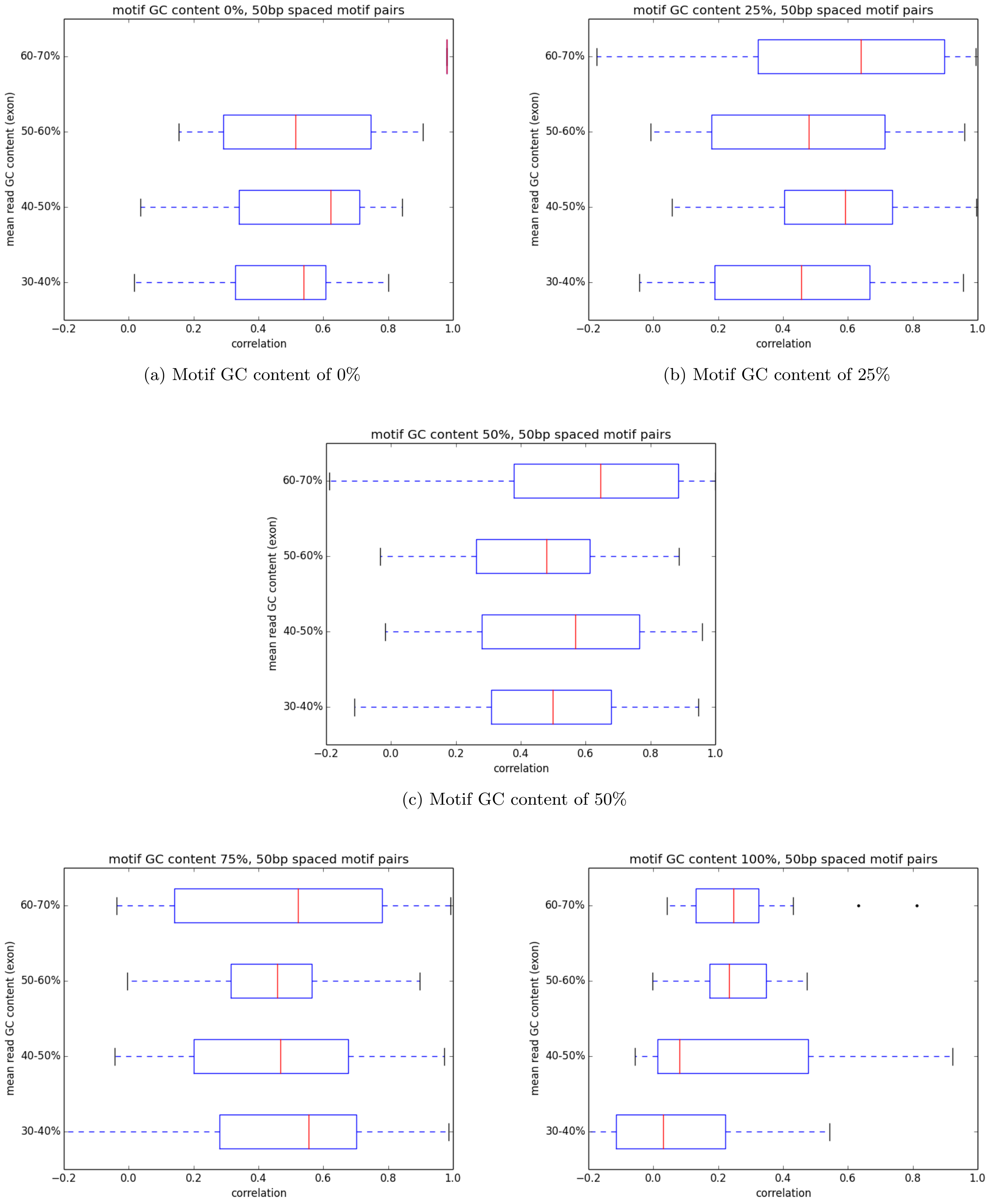
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.3: Box and whisker plots of motif-pair correlations at a distance of 50bp for Wild-type *D. melanogaster*

Wild type *D. melanogaster* - motif pair correlations at 50bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

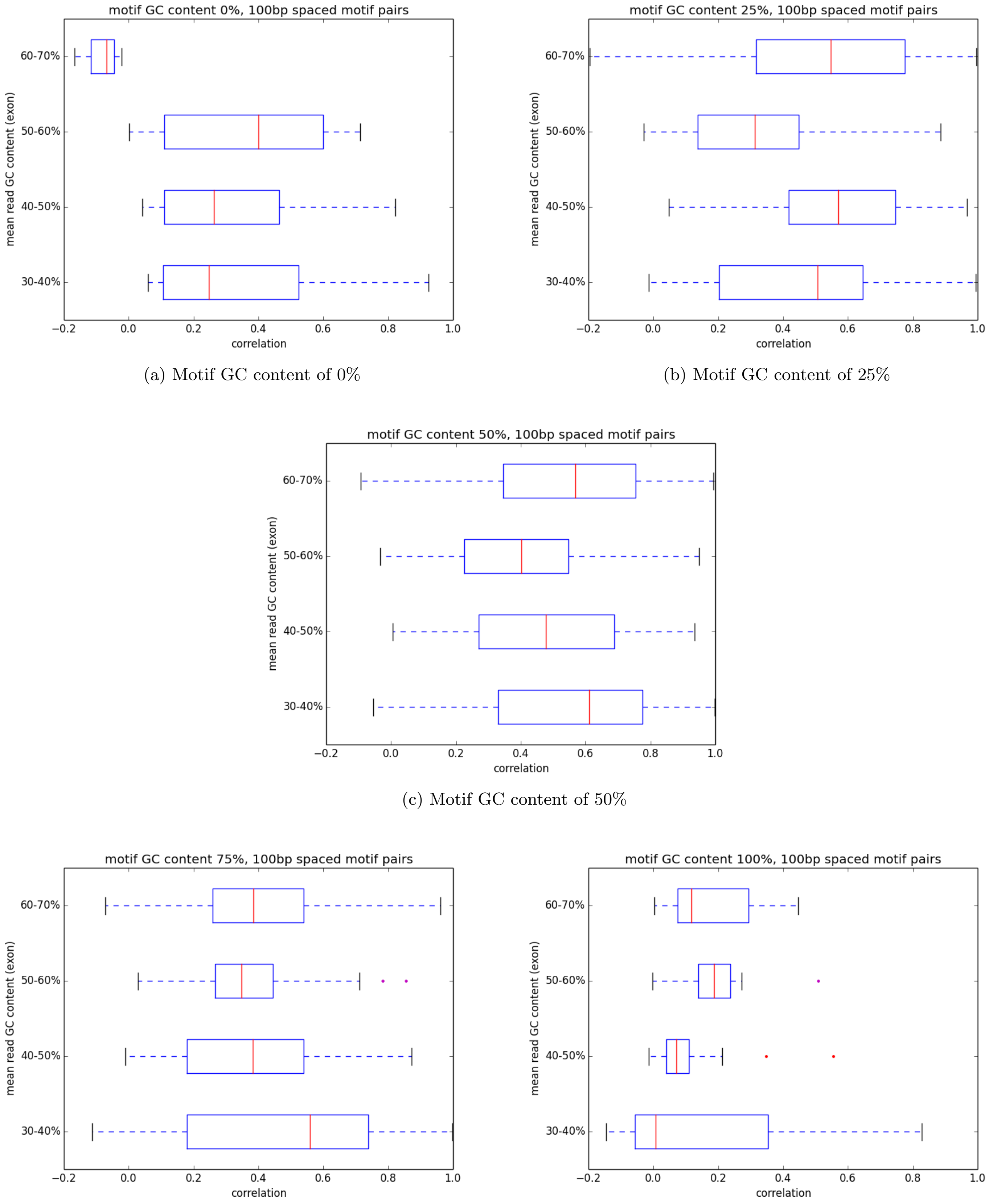
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.4: Box and whisker plots of motif pair correlations at a distance of 50bp for Wild-type *D. melanogaster* (Excluding hexamer regions)

Wild type *D. melanogaster* - motif-pair correlations at 100bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

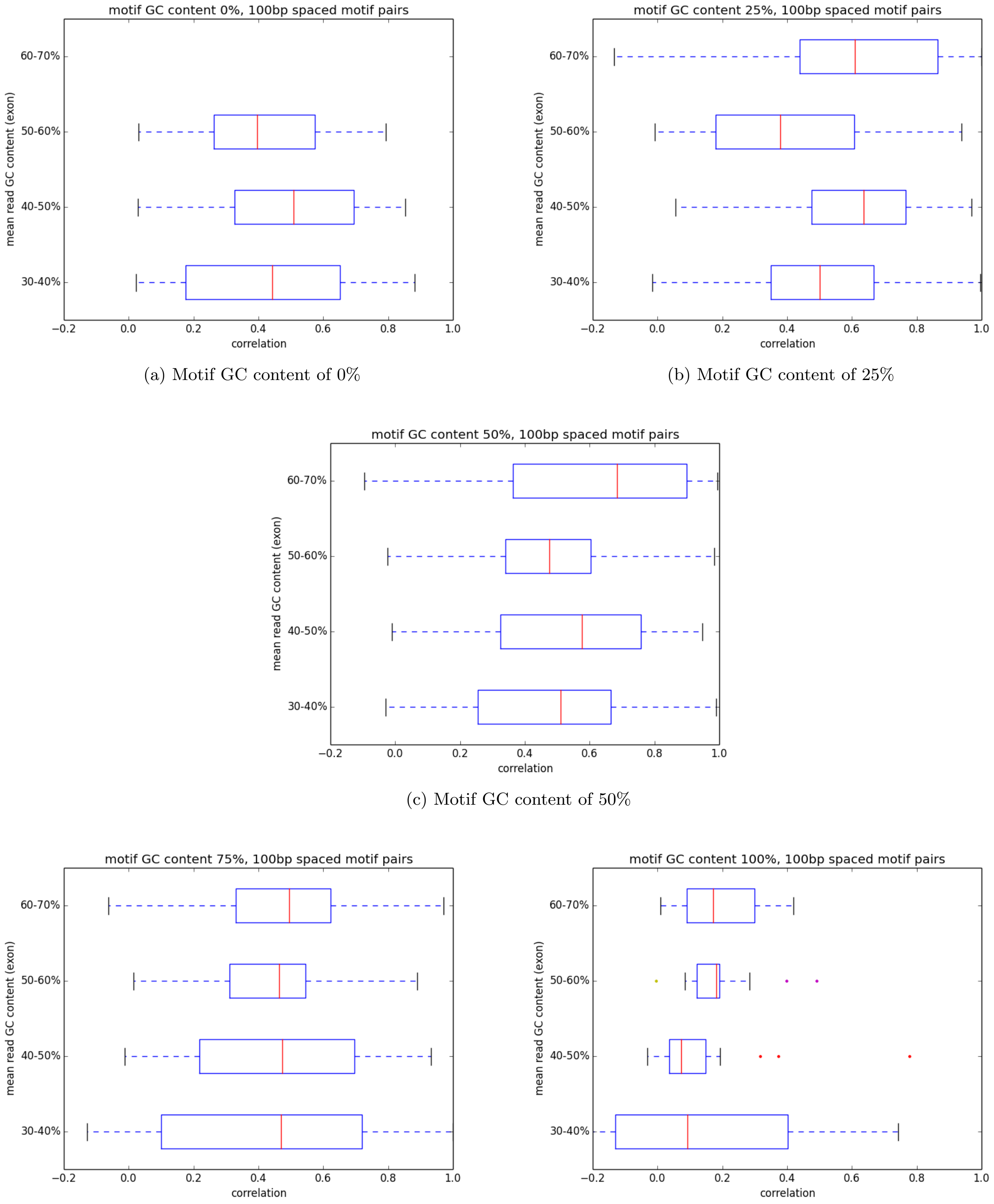
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.5: Box and whisker plots of motif-pair correlations at a distance of 100bp for Wild-type *D. melanogaster*.

Wild type *D. melanogaster* - motif pair correlations at 100bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.6: Box and whisker plots of motif pair correlations at a distance of 100bp for Wild-type *D. melanogaster* (Excluding hexamer regions)

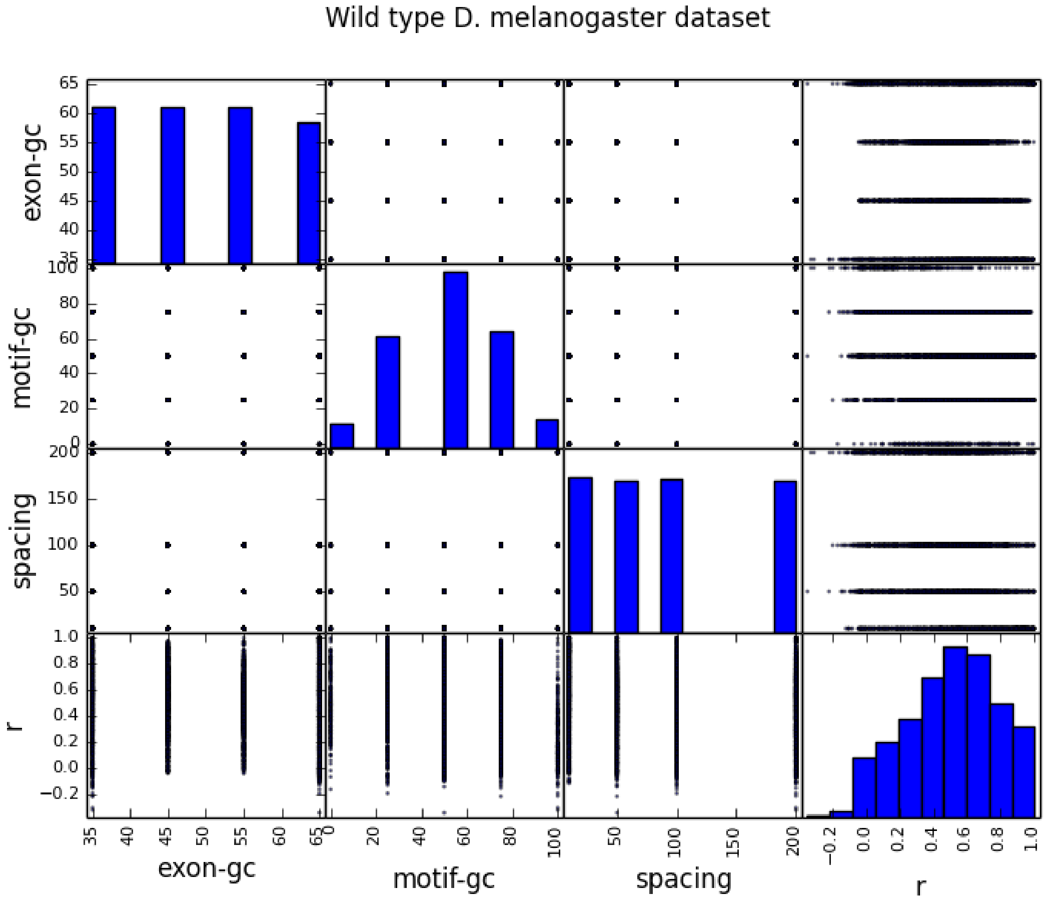


Figure 2.7: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in wild type *D. melanogaster*.

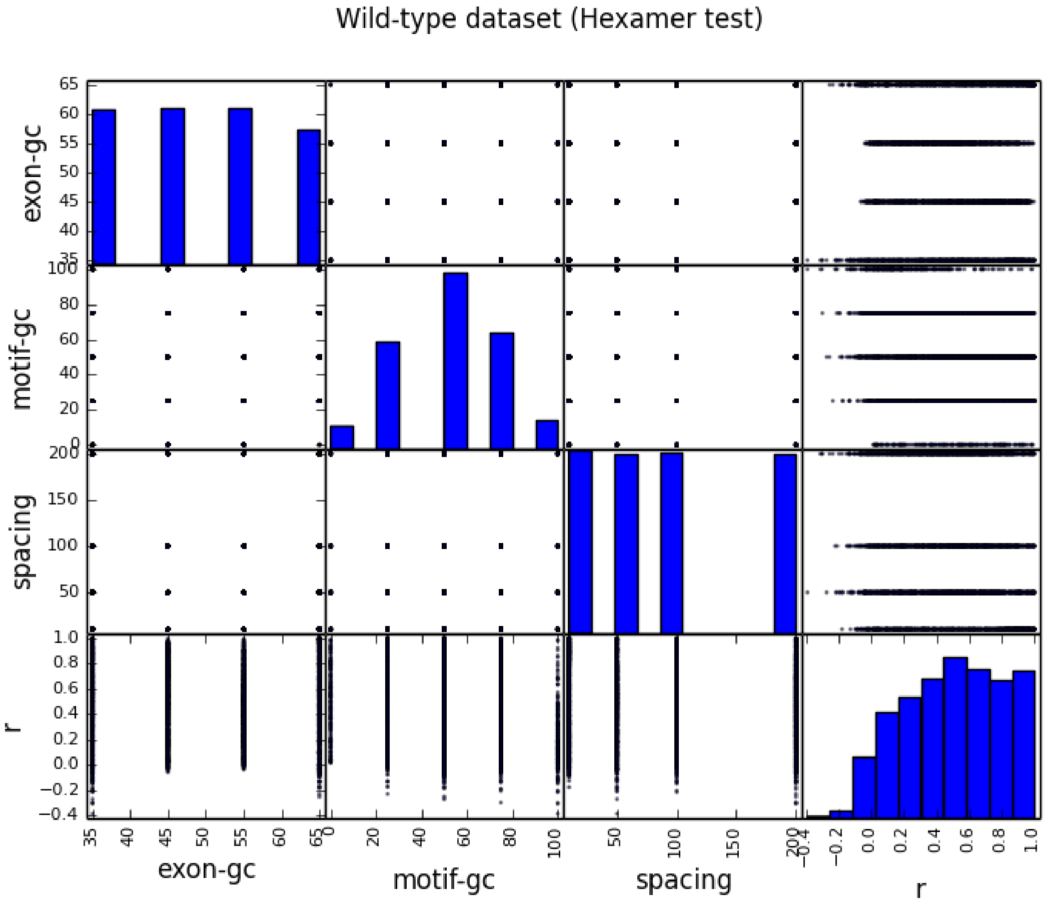


Figure 2.8: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in wild type *D. melanogaster*. Random hexamer priming region has been excluded.

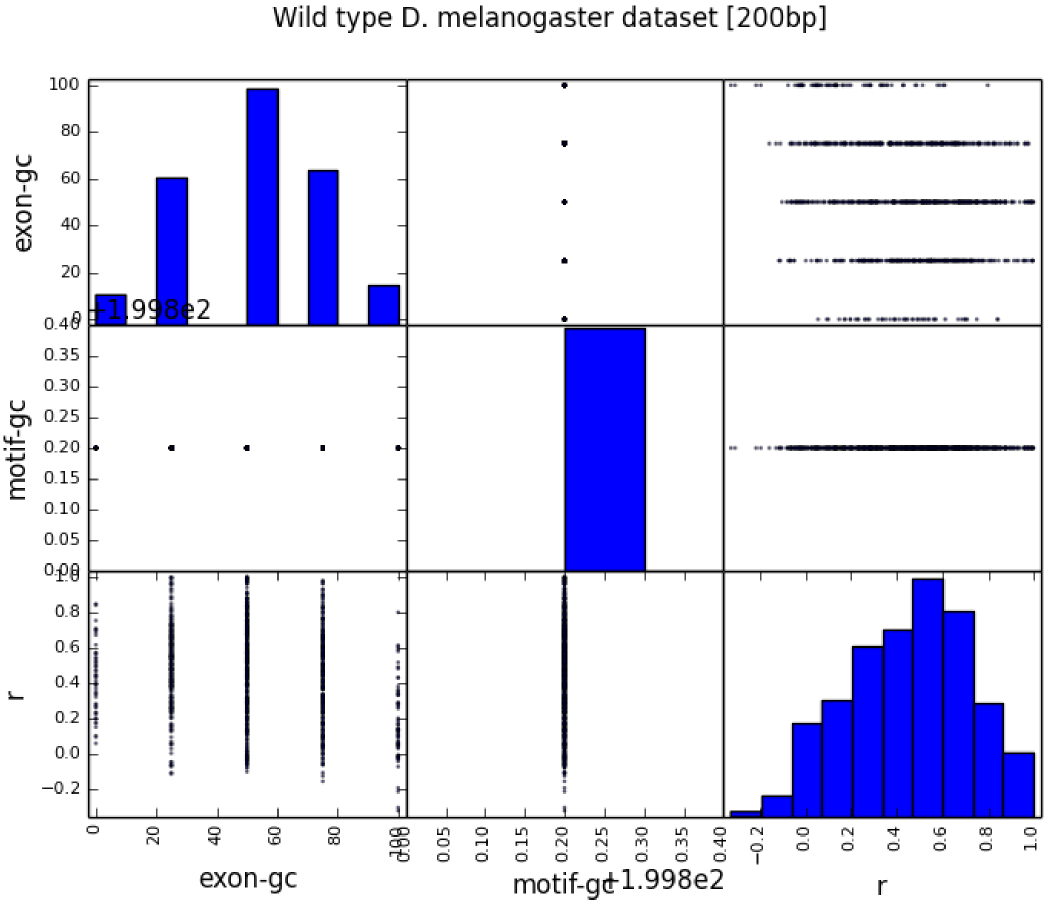


Figure 2.9: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in wild type *D. melanogaster* at a motif-pair spacing of 200bp.

2.2 Mutant-r2-type *D. melanogaster* results

Motif spacing: **10bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	$2.15\text{x}10^{-3}(1.15\text{x}10^{-2})$*		$3.94\text{x}10^{-1}(5.25\text{x}10^{-1})$		$2.02\text{x}10^{-1}(3.90\text{x}10^{-1})$		$9.55\text{x}10^{-1}(9.55\text{x}10^{-1})$	
p(Wilcoxon)	$6.05\text{x}10^{-1}(6.45\text{x}10^{-1})$		$5.69\text{x}10^{-1}(6.45\text{x}10^{-1})$		$3.26\text{x}10^{-1}(4.74\text{x}10^{-1})$		$1.48\text{x}10^{-1}(3.38\text{x}10^{-1})$	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	$2.19\text{x}10^{-1}(3.90\text{x}10^{-1})$		$9.51\text{x}10^{-1}(9.55\text{x}10^{-1})$		$5.28\text{x}10^{-2}(1.41\text{x}10^{-1})$		$2.93\text{x}10^{-1}(4.69\text{x}10^{-1})$	
p(Wilcoxon)	$3.09\text{x}10^{-1}(4.74\text{x}10^{-1})$		$5.74\text{x}10^{-1}(6.45\text{x}10^{-1})$		$5.25\text{x}10^{-2}(1.88\text{x}10^{-1})$		$2.83\text{x}10^{-2}(1.51\text{x}10^{-1})$	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	$3.26\text{x}10^{-1}(4.75\text{x}10^{-1})$		$8.44\text{x}10^{-1}(9.55\text{x}10^{-1})$		$4.35\text{x}10^{-3}(1.39\text{x}10^{-2})$*		$4.50\text{x}10^{-1}(5.54\text{x}10^{-1})$	
p(Wilcoxon)	$2.47\text{x}10^{-1}(4.40\text{x}10^{-1})$		$5.43\text{x}10^{-1}(6.45\text{x}10^{-1})$		$6.11\text{x}10^{-3}(9.77\text{x}10^{-2})$		$9.15\text{x}10^{-1}(9.15\text{x}10^{-1})$	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	$1.06\text{x}10^{-5}(1.69\text{x}10^{-4})$*		$3.16\text{x}10^{-3}(1.26\text{x}10^{-2})$*		$7.19\text{x}10^{-4}(5.75\text{x}10^{-3})$*		$6.88\text{x}10^{-2}(1.57\text{x}10^{-1})$	
p(Wilcoxon)	$2.62\text{x}10^{-2}(1.51\text{x}10^{-1})$		$7.03\text{x}10^{-2}(1.88\text{x}10^{-1})$		$6.27\text{x}10^{-2}(1.88\text{x}10^{-1})$		$1.79\text{x}10^{-1}(3.58\text{x}10^{-1})$	

Table 2.4: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 10bp spacing for varying motif GC and mean exon GC content in Mutant-r2 type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

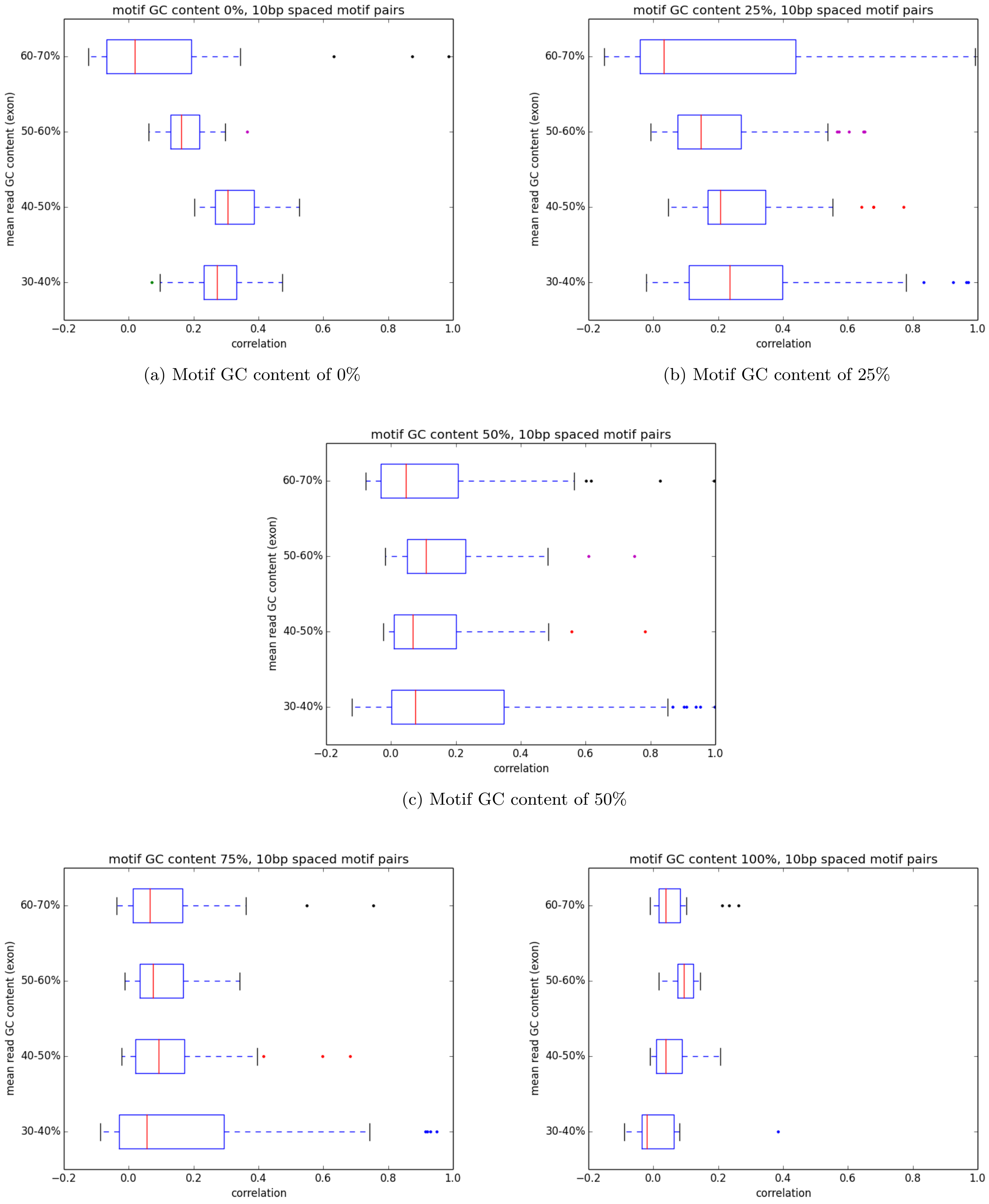
Motif spacing: **50bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	13	96
p(t-test)	9.09x10^{-1}(9.65x10^{-1})		8.87x10^{-1}(9.65x10^{-1})		2.40x10^{-1}(4.53x10^{-1})		9.65x10^{-1}(9.65x10^{-1})	
p(Wilcoxon)	8.79x10^{-2}(3.13x10^{-1})		5.35x10^{-1}(7.35x10^{-1})		1.79x10^{-1}(4.09x10^{-1})		8.07x10^{-1}(8.61x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	2.31x10^{-1}(4.53x10^{-1})		7.02x10^{-1}(8.99x10^{-1})		2.30x10^{-1}(4.53x10^{-1})		2.62x10^{-3}(2.02x10^{-2})*	
p(Wilcoxon)	4.85x10^{-2}(2.80x10^{-1})		2.11x10^{-1}(4.40x10^{-1})		1.11x10^{-1}(3.57x10^{-1})		1.63x10^{-3}(5.23x10^{-2})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	8.67x10^{-1}(9.65x10^{-1})		6.63x10^{-1}(8.85x10^{-1})		2.96x10^{-1}(4.99x10^{-1})		1.24x10^{-1}(3.98x10^{-1})	
p(Wilcoxon)	5.47x10^{-1}(7.35x10^{-1})		7.99x10^{-1}(8.61x10^{-1})		6.78x10^{-1}(7.75x10^{-1})		2.34x10^{-1}(4.40x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	4.76x10^{-2}(2.11x10^{-1})		6.06x10^{-1}(8.43x10^{-1})		1.42x10^{-1}(4.13x10^{-1})		2.07x10^{-1}(4.53x10^{-1})	
p(Wilcoxon)	6.42x10^{-1}(7.60x10^{-1})		9.59x10^{-1}(9.59x10^{-1})		1.63x10^{-1}(4.09x10^{-1})		2.34x10^{-1}(4.40x10^{-1})	

Table 2.5: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 50bp spacing for varying motif GC and mean exon GC content in Mutant-r2 type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

Motif spacing: **100bp**								
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	0	50	0	50	0	50	0
#Correlations	16	96	16	96	16	96	15	96
p(t-test)	7.05x10^{-1}(8.67x10^{-1})		1.84x10^{-1}(4.81x10^{-1})		4.86x10^{-1}(6.86x10^{-1})		2.26x10^{-1}(4.81x10^{-1})	
p(Wilcoxon)	7.56x10^{-1}(8.44x10^{-1})		1.63x10^{-1}(3.90x10^{-1})		4.69x10^{-1}(6.82x10^{-1})		1.06x10^{-2}(1.27x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	25	50	25	50	25	50	25
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	4.70x10^{-1}(6.86x10^{-1})		7.37x10^{-1}(8.85x10^{-1})		3.07x10^{-2}(1.47x10^{-1})		5.19x10^{-1}(7.12x10^{-1})	
p(Wilcoxon)	5.74x10^{-1}(7.07x10^{-1})		3.92x10^{-1}(6.07x10^{-1})		5.93x10^{-2}(2.73x10^{-1})		9.09x10^{-1}(9.34x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	75	50	75	50	75	50	75
#Correlations	64	96	64	96	64	96	64	96
p(t-test)	2.98x10^{-1}(5.10x10^{-1})		5.60x10^{-3}(3.36x10^{-2})*		4.78x10^{-1}(6.86x10^{-1})		9.54x10^{-1}(9.65x10^{-1})	
p(Wilcoxon)	2.83x10^{-2}(1.94x10^{-1})		8.69x10^{-2}(3.01x10^{-1})		3.00x10^{-1}(5.12x10^{-1})		3.00x10^{-1}(5.12x10^{-1})	
Exon GC%	**30-40%**		**40-50%**		**50-60%**		**60-70%**	
Motif GC%	50	100	50	100	50	100	50	100
#Correlations	16	96	16	96	16	96	16	96
p(t-test)	8.31x10^{-3}(4.43x10^{-2})*		1.86x10^{-5}(4.47x10^{-4})*		2.92x10^{-1}(5.10x10^{-1})		1.10x10^{-1}(3.76x10^{-1})	
p(Wilcoxon)	1.09x10^{-1}(3.34x10^{-1})		4.46x10^{-3}(9.77x10^{-2})		4.69x10^{-1}(6.82x10^{-1})		1.63x10^{-1}(3.90x10^{-1})	

Table 2.6: T-test and Wilcoxon-test comparisons of Pearson correlations for motif-pairs at 100bp spacing for varying motif GC and mean exon GC content in Mutant-r2 type *D. melanogaster*. FDR corrected p-values in parenthesis, using a False positive rate of 5% ($\alpha = 0.05$). * suggests rejection of the null hypothesis.

Mutant-r2 type *D. melanogaster* - motif-pair correlations at 10bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

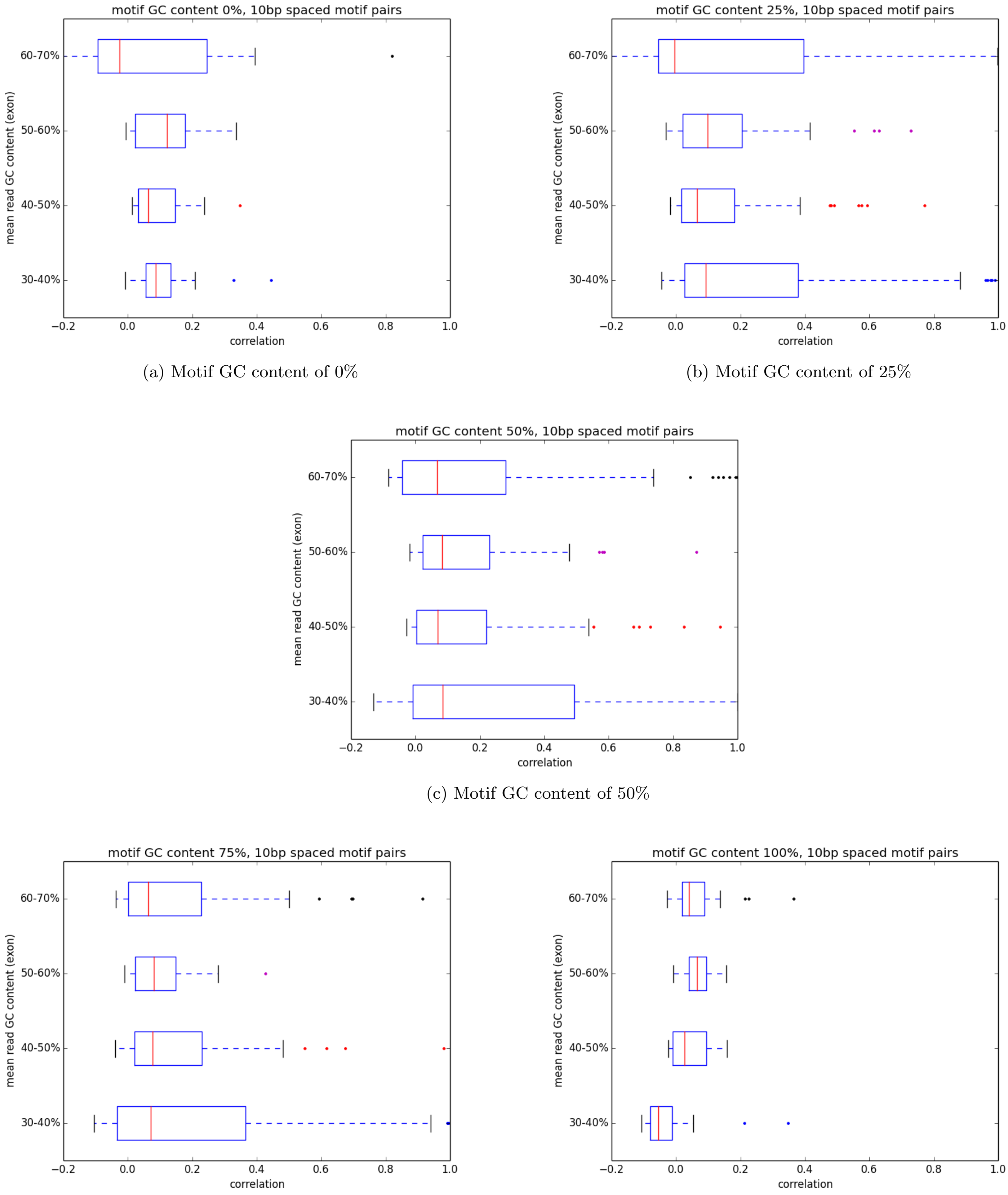
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.10: Box and whisker plots of motif-pair correlations at a distance of 10bp for Mutant-r2-type *D. melanogaster*

Mutant-r2 type *D. melanogaster* - motif pair correlations at 10bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

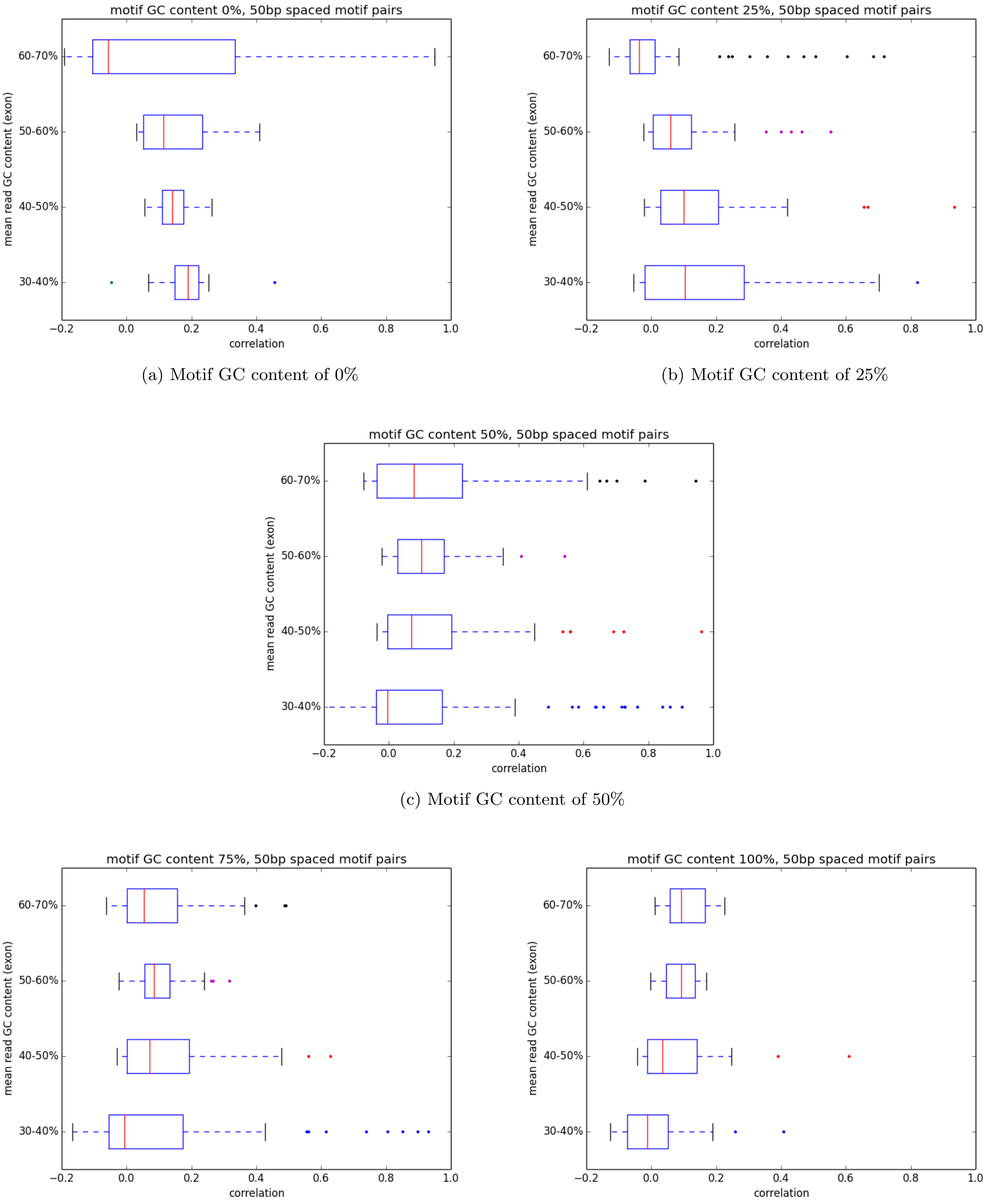
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.11: Box and whisker plots of motif pair correlations at a distance of 10bp for Mutant-r2-type *D. melanogaster* (Excluding hexamer regions)

Mutant-r2 type *D. melanogaster* - motif-pair correlations at 50bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

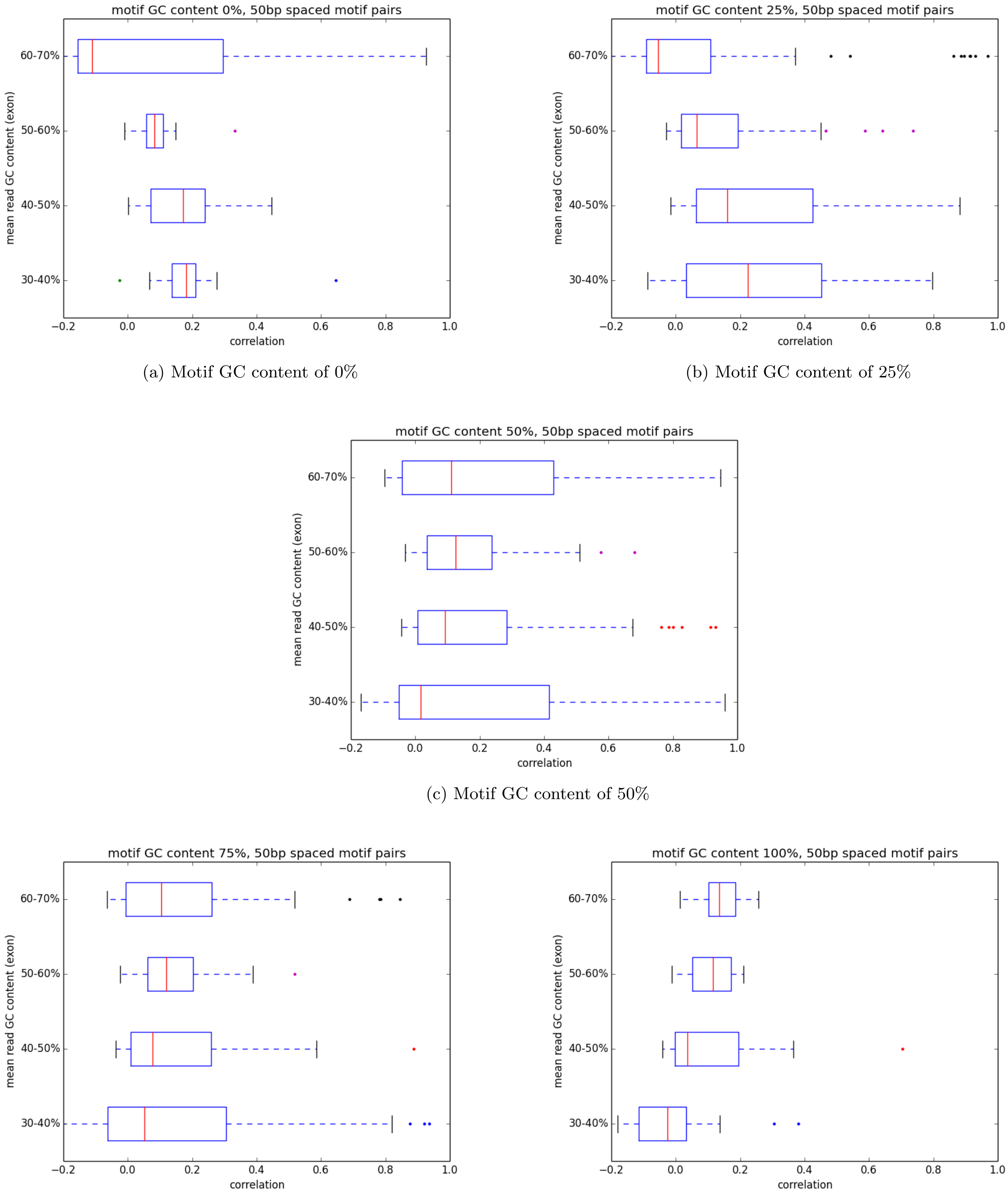
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.12: Box and whisker plots of motif-pair correlations at a distance of 50bp for Mutant-r2-type *D. melanogaster*

Mutant-r2 type *D. melanogaster* - motif pair correlations at 50bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

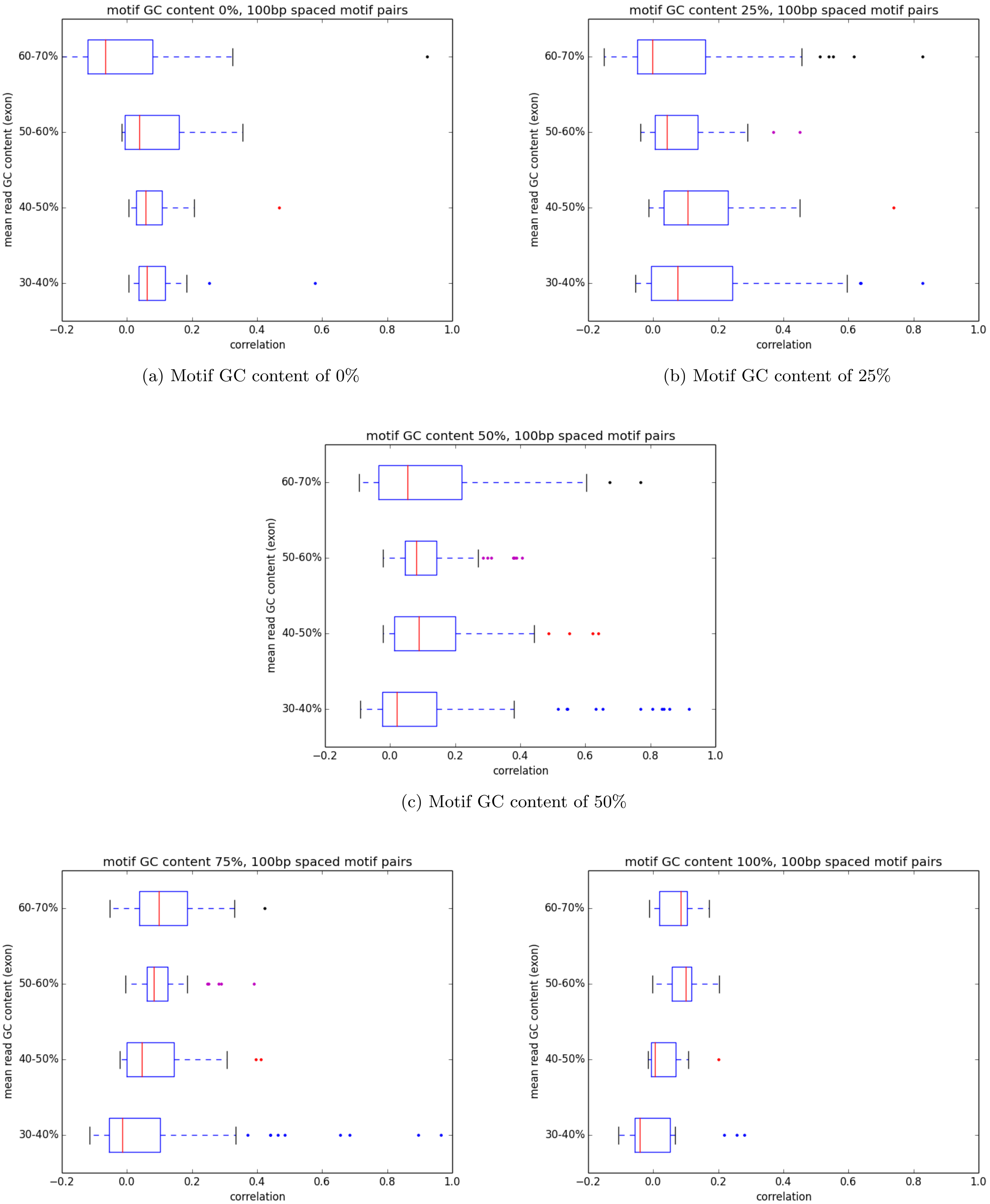
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.13: Box and whisker plots of motif pair correlations at a distance of 50bp for Mutant-r2-type *D. melanogaster* (Excluding hexamer regions)

Mutant-r2 type *D. melanogaster* - motif-pair correlations at 100bp apart



(a) Motif GC content of 0%

(b) Motif GC content of 25%

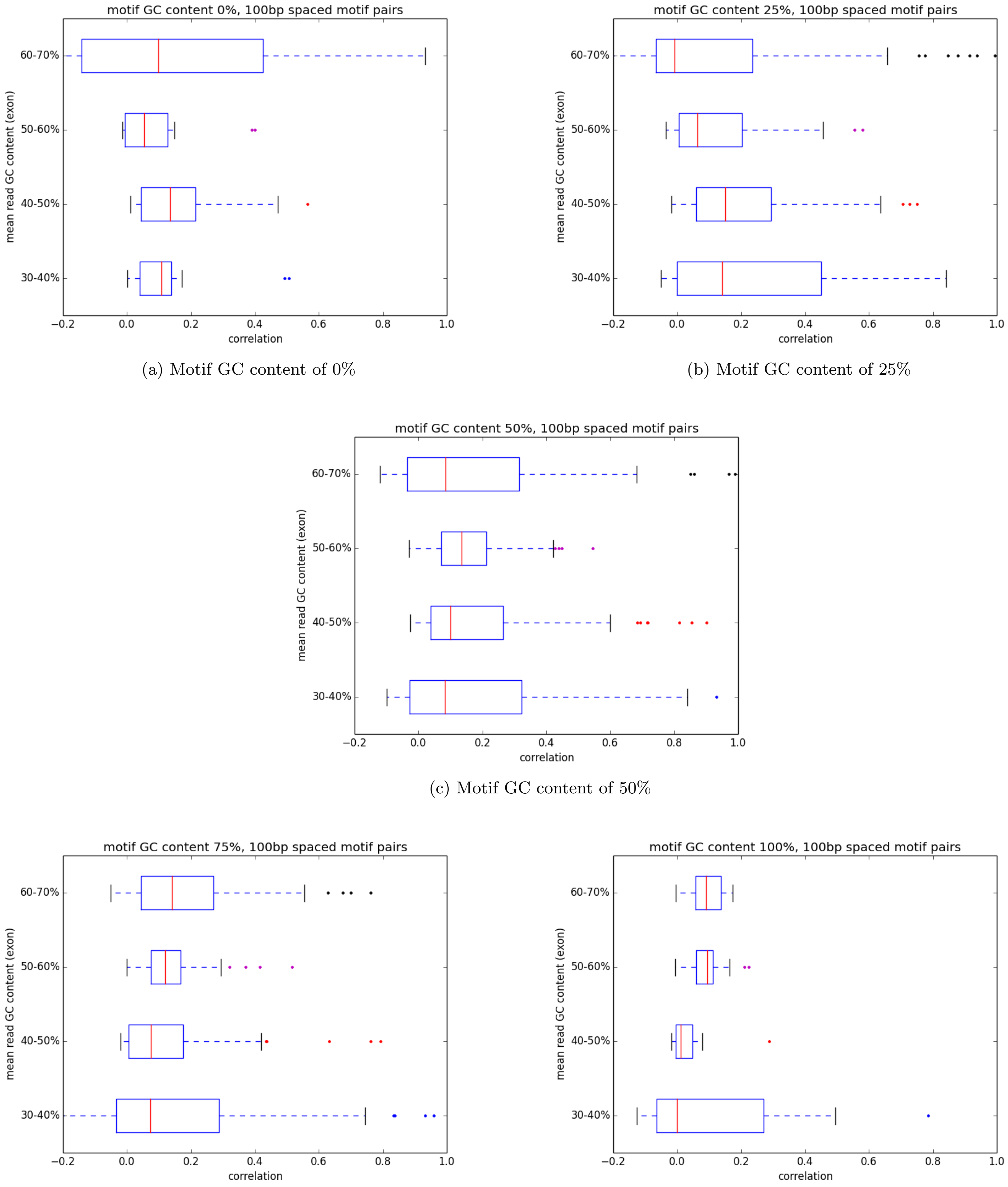
(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.14: Box and whisker plots of motif-pair correlations at a distance of 100bp for Mutant-r2-type *D. melanogaster*.

Mutant-r2 type *D. melanogaster* - motif pair correlations at 100bp apart (excluding hexamer primers)



(a) Motif GC content of 0%

(b) Motif GC content of 25%

(c) Motif GC content of 50%

(d) Motif GC content of 75%

(e) Motif GC content of 100%

Figure 2.15: Box and whisker plots of motif pair correlations at a distance of 100bp for Mutant-r2-type *D. melanogaster* (Excluding hexamer regions)

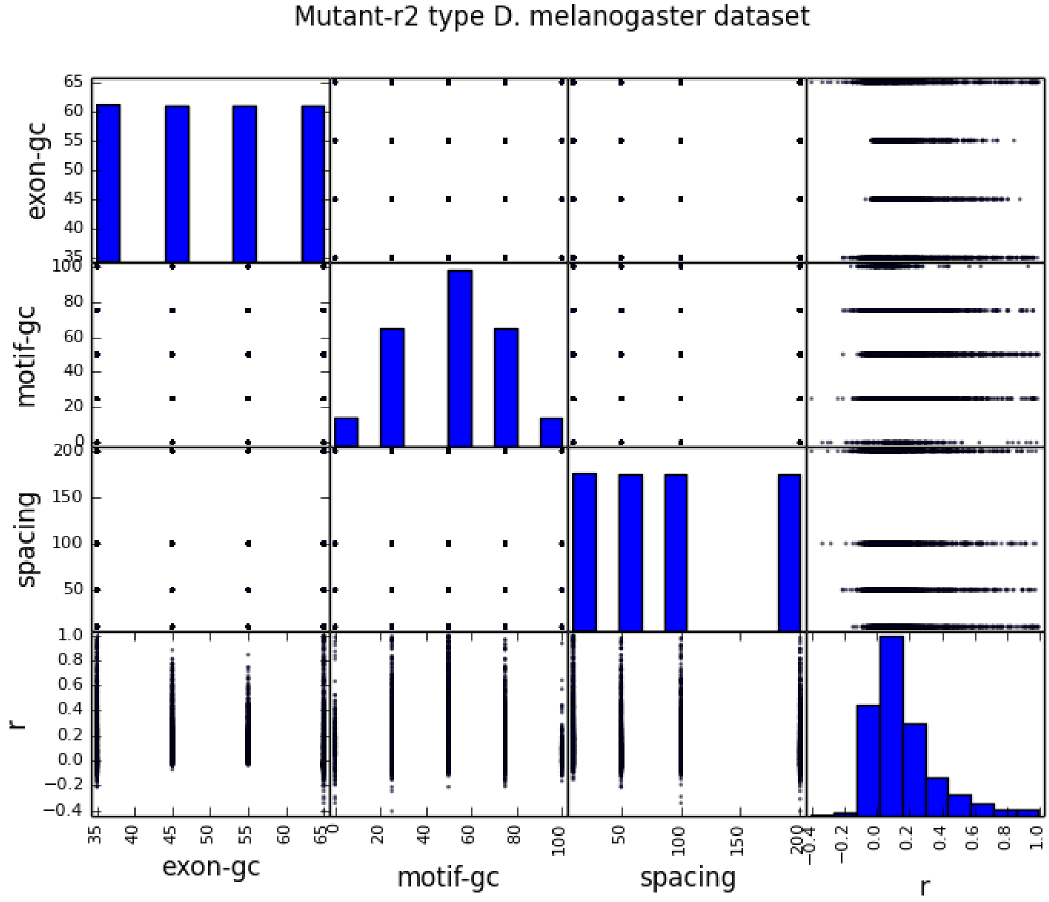


Figure 2.16: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in mutant-r2 *D. melanogaster*.

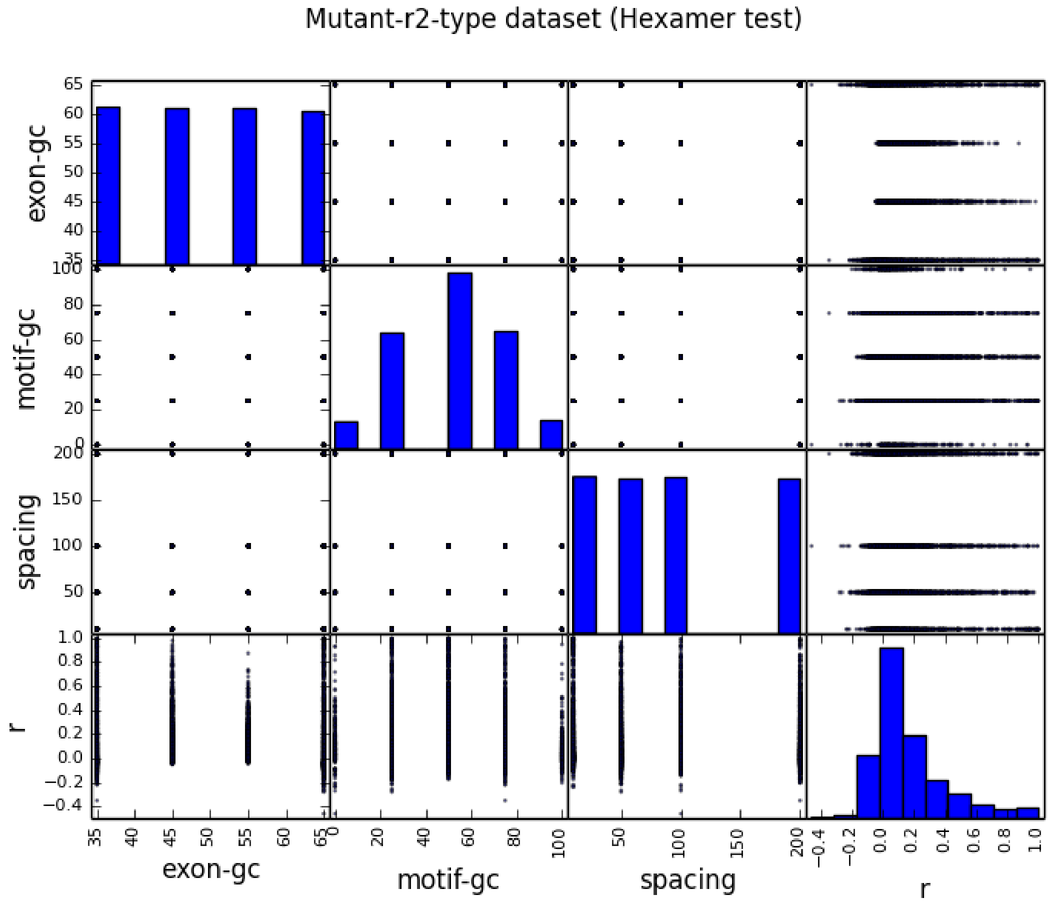


Figure 2.17: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in mutant-r2 *D. melanogaster*. Random hexamer priming region has been excluded.

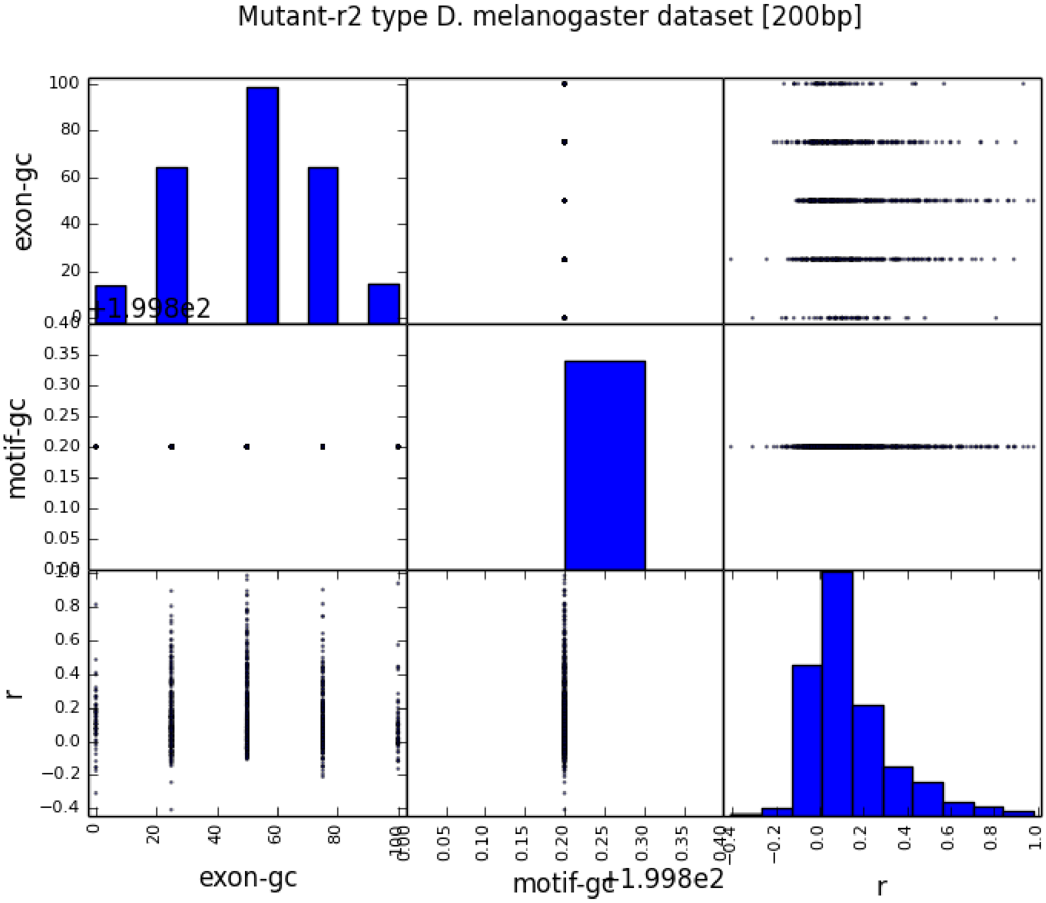


Figure 2.18: Scatter-matrix plot of correlation as a function of *4-mer* motif and exon GC content in mutant-r2 *D. melanogaster* at a motif-pair spacing of 200bp.